

基于大型多模态模型的街景图像典型场景要素提取

潘可悦¹, 芮维¹, 程湘¹, 刘异²

(1. 武汉大学测绘遥感信息工程国家重点实验室, 湖北 武汉 430079; 2. 武汉大学测绘学院, 湖北 武汉 430079)

摘要: 场景要素是理解城市地理信息的核心, 准确提取场景要素对于智慧城市建设和地理信息系统发展至关重要。为应对街景图像场景的复杂性、现有视觉深度学习模型在理解复杂场景和要素方面的局限性, 以及视觉信息与上下文关联的挑战, 提出了一种基于大型多模态模型(LMM)的典型街景场景要素提取方法。首先, 基于 LLaVA (Large Language and Vision Assistant) 模型引入多层感知机 (MLP) 和高分辨率视觉编码器, 构建 GeoLLaVA 模型; 其次, 针对街景场景要素提取任务构建街景视觉-指令跟随数据集, 提供多维度指令, 通过视觉指令微调模型, 增强其对复杂街景场景的上下文理解, 同时, 引入低秩自适应 (LoRA) 技术降低计算资源需求; 最后, 通过 GeoLLaVA 模型生成街景图像的多维度场景描述, 并提取关键词以获得典型场景要素。在与语义分割、目标检测及其他多模态模型的对比实验中, GeoLLaVA 表现出了显著优势, 在交通信号灯、交叉路口和停车场要素提取任务中分别取得了 0.938、0.842 和 0.829 的 F1 值。模型微调前后的对比展现了微调的有效性。消融实验进一步验证了 GeoLLaVA 改进结构对性能提升的贡献以及 LoRA 在降低计算资源消耗方面的有效性。区域应用实验通过批量推理特定区域的街景图像, 提取要素并结合地理位置进行可视化展示, 与开放街景地图 (OSM) 数据对比, 验证了模型的准确性并揭示了 OSM 在提供要素信息方面的不足。

关键词: 大型多模态模型; 典型场景要素; LLaVA 模型; 街景图像; 低秩自适应

源代码链接: <https://github.com/Smerfy666/GeoLLaVA1.0>

中图分类号: TP391

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0070469

Extraction of Typical Scene Elements from Street View Images Based on Large Multimodal Models

PAN Keyue¹, GUO Wei¹, CHENG Xiang¹, LIU Yi²

(1. State Key Laboratory of Information Engineering in Surveying,

Mapping and Remote Sensing, Wuhan University, Wuhan 430079, Hubei, China;

2. School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, Hubei, China)

【Abstract】 Scene elements are fundamental for understanding urban geographic information, and their accurate extraction is essential for smart city development and geographic information systems. To address the complexity of street-view images, the limitations of existing deep learning models in interpreting complex scenes, and challenges in associating visual data with context, a method based on large multimodal models for extracting typical scene elements from street-view images is proposed. First, the approach extends the Large Language and Vision Assistant (LLaVA) by integrating a multilayer perceptron and a high-resolution visual encoder to create GeoLLaVA. Second, a Street View Visual-Instruction-Following Dataset is constructed for scene element extraction tasks to provide multidimensional instructions. The model is fine-tuned using visual instructions to enhance contextual understanding. Low-Rank Adaptation (LoRA) is used to optimize the computational efficiency. Finally, GeoLLaVA generates multidimensional scene descriptions from street-view images and extracts key element keywords for effective scene-element extraction. In comparative experiments using semantic segmentation, object detection, and other multimodal models, GeoLLaVA demonstrates significant advantages, achieving F1 scores of 0.938, 0.842, and 0.829 for the extraction of traffic signals, intersections, and parking lots, respectively. A comparison between the model before and after fine-tuning clearly demonstrates the effectiveness of the fine-tuning process. Ablation studies further validate the performance improvements achieved by the modified GeoLLaVA architecture, and LoRA effectively reduces computational resource consumption. Regional application experiments using batch inference on street view images with geographic coordinates and a comparison with OpenStreetMap (OSM) data not only confirm the accuracy of the model but also highlight the limitations of OSM data in providing comprehensive element information.

基金项目: 国家自然科学基金(42071431)。

作者简介: 潘可悦, 女, 硕士研究生, 主研方向为大模型技术、视觉地理定位算法、遥感影像数据处理技术; 芮维, 教授; 程湘 (通信作者), 硕士研究生; 刘异, 副教授。

收稿日期: 2024-10-11

修回日期: 2025-01-29

E-mail: xiang_cheng@whu.edu.cn

【Key words】 Large Multimodal Model (LMM); typical scene elements; Large Language and Vision Assistant (LLaVA); street view images; Low-Rank Adaptation (LoRA)

0 引言

随着城市化进程加快和智慧城市建设推进,城市环境中的地理信息已然成为城市规划、交通管理等领域的关键需求。建筑物、道路、交通设施等典型场景要素相互交织,共同塑造着区域的地理特征与功能布局。理解这些要素的分布及相互关系,有助于深度挖掘地理信息,为区域规划、交通管理等工作提供有力的数据支撑。例如,街道上的交通信号灯、路口、停车场等要素的分布,成为智慧交通建设的关注焦点。由此可见,典型场景要素的准确提取对于城市规划、交通系统的高效运作以及地理信息系统(GIS)的发展具有重要意义^[1]。然而,此过程面临数据与技术双重困境。

首先,数据层面的局限性尤为突出。以开放街道地图(OSM)为代表的平台虽然提供了丰富的地理要素信息,但其数据依赖人工收集,存在准确性低、完整性差、更新滞后等问题,削弱了其应用价值。近年来,街景影像凭借丰富的地理信息和空间细节,成为城市分析和地理空间数据收集的重要视觉资源^[2],基于街景影像提取场景要素已成为新的研究趋势。然而,街景图像场景复杂且多样化,地理元素交错,区域差异明显,且容易受光照、视角等因素影响,这使得基于街景图像的要素提取并不容易。

其次,在技术层面,街景图像场景的复杂性对现有方法在要素提取时造成一系列的挑战。传统计算机视觉方法通过特征工程来应对图像复杂性^[3],但手工特征难以全面捕捉场景信息。深度学习兴起后,基于卷积神经网络等架构的深度学习模型被用于街景要素提取^[4],常见方法有语义分割^[5]和目标检测^[6],一定程度上增强了提取效果。但这些方法仅依赖图像视觉信息,能识别如交通信号灯、行人、车辆等视觉特征明显的要素^[7-8],但在理解复杂的地理要素间空间关系和适应多样化的复杂场景(如交叉路口和停车场)时,相关研究尚显不足,难以提供有效的解决方案。在此背景下,研究者意识到单一模态信息的局限,进而开展多模态学习方法的研究^[9]。多模态学习指通过融合多种模态的数据(如文本、图像、音频等)来训练模型,从而增强模型的性能和理解能力^[10]。近年来,小型多模态模型应用广泛,但因数据稀缺和信息抽象等问题,需大量外部知识补充及结构的精细优化,限制了其泛化能力^[11]。

另外,小型多模态模型虽尝试融合视觉与其他模态信息,但在挖掘视觉信息与地理环境及功能布局的上下文关联方面仍显欠缺。以 CLIP(Contrastive Language-Image Pre-training)模型为例,它虽能识别交通标志和建筑物类别,却对这些要素与周围环境的关联及其在街景场景中的作用理解深度有限^[12]。然而,大型多模态模型(LMM),如 GPT-4(Generative Pre-trained Transformer 4)^[13] 和 LLaVA(Large Language and Vision Assistant)^[14],通过精心的提示设计与大规模、复杂的预训练,能够基于上下文信息提供更丰富、针对性更强的分析。这使 LMM 在面对高度复杂且多层次的场景时,表现出更强的泛化能力与更深的上下文理解能力^[15-16],在应对街景场景要素提取难题方面展现出显著潜力。

鉴于此,本文提出了一种基于大型多模态模型的街景图像典型场景要素提取方法,构建基于预训练大型语言模型^[17]和多模态视觉模型^[18]的 GeoLLaVA 模型。该模型在 LLaVA 基础上进行创新,引入多层感知机(MLP)跨模态连接器并采用高分辨率视觉编码器,采用 Transformer 等先进的深度神经网络架构与大规模预训练知识,通过跨模态学习融合视觉与语言信息,精准解析语义关系,提取复杂街景要素。为了引导模型聚焦于场景要素提取相关的视觉和上下文信息,本文构建了一个街景视觉-指令跟随数据集,并通过视觉指令微调方法提升模型性能。然而,大型模型的微调面临一系列挑战:模型参数庞大,进行全参数微调需要大量计算资源,且有限的训练数据可能导致过拟合,影响模型的泛化能力。为此,本文引入低秩自适应(LoRA)^[19]技术,通过在关键层引入低秩矩阵来调整权重,减少需要更新的参数数量,降低对计算资源的需求,并有效减少过拟合风险。在保持模型预训练知识的同时,LoRA 策略能够有效地微调模型,使其更好地理解 and 处理街景图像中的视觉信息及上下文关系。

本文提出的方法充分发挥了 LMM 在融合视觉和语言信息方面的优势,有效应对了街景图像场景的复杂性、现有的基于视觉信息的单模态深度学习模型在理解复杂场景和要素方面的局限性,以及视觉信息与上下文关联的挑战。该方法通过整合街景图像中的视觉信息与地理环境背景,精确提取街景典型场景要素,为相关领域提供更加准确、全面的数

据支持,不仅为多模态大模型在实际应用中的优化提供了新思路,同时推动了 GIS 的自动化与智能化发展,为城市规划与管理提供了创新解决方案,具有重要的应用价值与深远的研究意义。

本文的主要贡献总结如下:

1) 模型架构创新。基于 LLaVA 构建 GeoLLaVA 模型,引入 MLP 跨模态连接器和高分辨率视觉编码器,增强模型对视觉与语言的感知与融合能力。

2) 构建街景视觉-指令跟随数据集。本文针对街景要素提取任务,精心设计提示指令库,并利用 GPT-4 辅助生成的视觉-指令跟随数据,以增强模型对街景场景分析相关指令的理解和执行能力。

3) 场景要素提取流程与应用创新。本文将大型多模态模型用于街景要素提取,利用构建的数据集与 LoRA 策略微调模型,精准提取街景要素,并结合地理信息实现要素在城市空间的可视化展示与深度分析。

1 相关研究

街景场景要素提取是地理信息理解与描述的关键,随着城市化进程推进,其在地理信息管理中的作用日益重要。然而,该技术面临诸多挑战,亟须探索有效方法以提升模型在街景场景要素提取任务中的表现。

1.1 传统计算机视觉方法

早期街景图像分析依赖传统计算机视觉方法,多运用浅层、中层及人工设计的特征来获取图像信息,借助如尺度不变特征变换(SIFT)、方向梯度直方图(HOG)等特征描述器进行边缘检测、形态分析、角点检测等操作,以提取图像的边缘、形状、颜色等特征^[20],并与机器学习算法相结合构建图像处理任务。例如,文献[3]利用 HOG 结合支持向量机(SVM)分类器,从街景中提取能代表巴黎城市风格的建筑元素;文献[21]利用边缘检测算法与 SIFT 特征来提取建筑物的轮廓。但这些方法依赖大量的专家知识和时间进行特征工程,且低层特征对视角、光照等因素敏感,难以完全捕捉复杂语义信息 and 理解复杂的街景场景,泛化能力较差。

1.2 深度学习方法

深度学习方法的的发展,为挖掘街景图像语义信息、理解和表达场景物质空间元素提供了有力支撑^[22]。目标检测、语义分割等方法被广泛应用于街景要素提取。例如,文献[23]运用基于单一前馈卷积神经网络的 SSD MobileNet 模型检测交通标志;

文献[24]采用 PSPNet^[7]提取街景像素,借助金字塔池化模块提供的额外上下文信息避免错误分割;文献[8]使用 DeepLabV3+^[25]对谷歌街景(GSV)图像执行语义分割以获取不同类别的街景要素,利用空洞空间金字塔池化方法高效提取多尺度语义信息;文献[26]则是集成了 HAMM(Hierarchical Adaptive Mean Shift)、DeepLabV3 和 YOLACT(You Only Look at Coefficients)对人行道进行提取与检测;文献[6]利用 YOLO(You Only Look Once)模型^[27]来识别夜间的道路标志。尽管这些方法取得了显著成果,但仍存在局限性。一方面,这些模型通常在特定的数据集(Cityscape^[28]、ADE20K^[29]等)上训练,涵盖的街景场景和元素有限,面对新的街景要素时模型的泛化能力较弱;另一方面,街景场景往往具有多层次、复杂的空间关系,单模态模型局限于从单一的视觉角度进行特征提取,忽略了元素间更为复杂的语义关联和全局上下文,限制了其在实际复杂街景场景中的应用价值。

1.3 多模态学习方法

当下,融合视觉-语言的多模态学习方法为街景场景理解和要素提取开辟了新的路径。这些多模态模型不仅能够识别图像中的各类要素和属性,还能通过语言上下文的补充捕捉元素之间的复杂关系和全局背景,进而提升模型的泛化能力和理解深度。例如,文献[30]使用视觉-语言模型,结合自底向上和自顶向下的注意力机制进行训练,以生成精确和流畅的街景描述;CLIP 模型^[18]的提出标志着多模态预训练模型的革命,它通过大规模图像-文本对比学习,理解两者之间的语义关系,具备较强的泛化能力;文献[12]通过微调 CLIP 模型,结合健康相关的图像-文本数据,提取街景场景中与健康相关的要素(如绿地、运动场)以预测城市的抑郁率;文献[31]同样利用了 CLIP,将街景图像和 OSM 土地利用上下文作为输入,以预测图像地理位置的主要土地利用类型。尽管这些多模态模型在提取高层次语义信息方面取得了一定成果,但面对地理信息中高度复杂或多层次的空间结构时,其泛化能力和上下文理解能力仍需进一步提升。

1.4 大型多模态模型

为提升多模态模型在深层次理解、内容生成及广泛应用方面的能力,大型多模态模型顺势而生。这些模型通过整合丰富的数据资源、优化架构和采用先进训练算法,推动了多模态理解的进步。其中,视觉-语言大模型能挖掘图像场景中深层次信息并生成富含语义的场景描述,在医学^[32]、工业^[33]等领

域发展迅速。然而, LMM 在地理信息领域的应用至今仍然相对有限。文献[34]中基于 GPT-4 开发的 GeoLocator 可从图像推断位置信息, 其核心是强调关注图像细节的明确指令; 在街景场景的应用中, 文献[35]使用 ChatGPT 为灾害前后的街景图片生成文本标题, 通过补充图像视觉信息提升建筑损坏评估的准确性和全面性; 文献[36]利用 ShareGPT4V 为街景图像生成文本描述, 通过设计包含分割元素比例和地理空间坐标的提示模板来引导表达。尽管这些研究开始采用大型多模态模型为街景图像提供描述, 但尚未对场景要素提取任务进行针对性的模型优化。分析其原因, 一方面, 用于街景场景要素提取的数据集有限, 难以满足模型深入学习和优化的需求; 另一方面, 大型模型参数量大, 全参数微调耗费资源且易过拟合, 影响模型在新任务上的泛化能力。对此, LoRA^[19]作为一种高效的

微调策略, 提供了一种有效的解决方案。这一技术的引入, 为大型多模态模型在复杂街景场景要素提取中的应用提供了新的突破口。

2 方法

2.1 模型架构

本文构建的大型视觉-语言模型结构如图 1 所示(彩色效果见《计算机工程》官网 HTML 版, 下同)。本文以 LLaVA 模型为基线模型, 利用视觉编码器提取输入图像的视觉特征, 经过单层线性映射层映射到语言空间, 从而实现视觉与语言信息的融合。随后, 预训练的大型语言模型 Vicuna^[17]将对这些融合后的特征进行理解和生成。为进一步提升模型性能, 本文做了两方面的改进, 其一是使用更高分辨率的视觉编码器, 其二是引入 MLP 跨模态连接器。

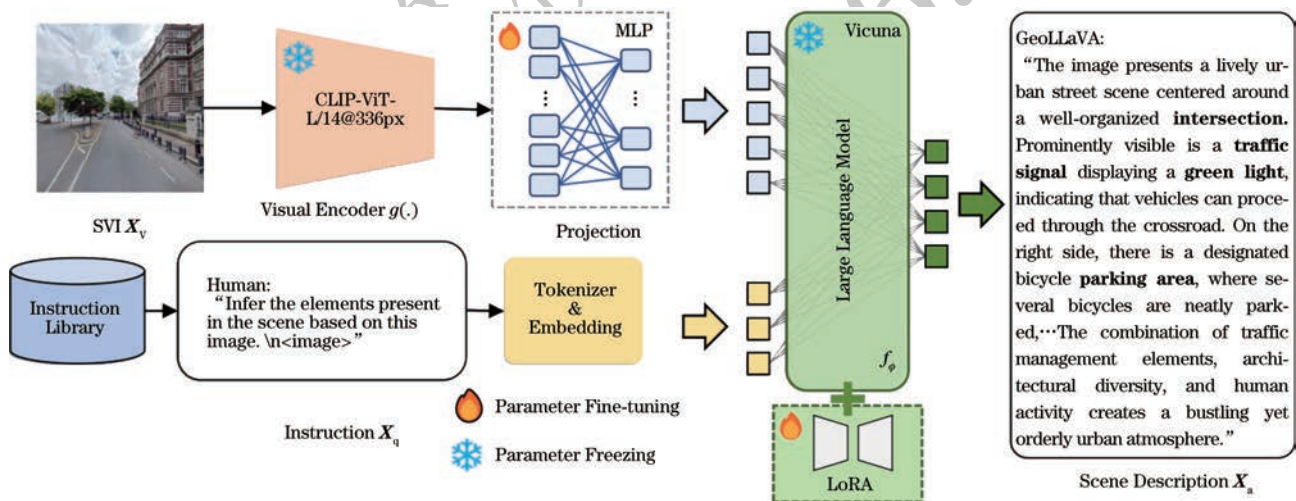


图 1 GeoLLaVA 模型框架图

Fig.1 GeoLLaVA architecture diagram

2.1.1 视觉编码器

本文所采用的视觉编码器来自 CLIP 模型的优秀变种: CLIP-ViT-L/14-336px。CLIP 模型是由文本编码器和图像编码器组成的多模态预训练模型, 该模型通过在 4 亿对图像-文本数据上进行对比学习, 具备了强大的泛化能力, 其图像编码器基于 ViT (Vision Transformer) 架构^[37], 通过 Transformer 的自注意力机制有效提取图像的全局特征。与 LLaVA 中使用的 CLIP-ViT-L/14 相比, CLIP-ViT-L/14-336px 将输入图像的分辨率提升至 336×336 像素, 分辨率的提升使其能提取到更丰富的视觉细节, 帮助模型更精准地“感知”高清街景图像的内容。

具体而言, CLIP-ViT-L/14-336px 首先将输入图像分割为多个不重叠的图像块, 各图像块经线性

嵌入映射至高维空间。然后, ViT 运用 Transformer 自注意力机制处理图像块, 捕捉区域依赖关系以提取全局特征。最后, 图像特征被映射为固定维度的特征向量, 用于后续文本与视觉信息融合处理。这一过程可简单表示为式(1):

$$Z_V = g(X_V) \quad (1)$$

式中: X_V 表示输入的图像; $g(X_V)$ 表示视觉编码器提取的视觉特征; Z_V 为生成的图像特征向量。

2.1.2 跨模态连接器

由于直接提取的视觉特征和由提示指令生成的词向量处于不同维度空间, 因此需要一个高效的连接器来实现视觉特征到与语言嵌入相同的特征空间的投影, 以实现两者的精准对齐与连接。受文献[38]将线性投影改进为 MLP 实现性能提升的启发, 本文引入一个可训练的 MLP^[39] 作为连接器, 取代原

始 LLaVA 模型中的单一线性层。MLP 通过引入非线性激活函数和隐藏层,有效地捕捉复杂的映射关系,增强了模型对视觉与语言信息的融合能力。

在具体运行过程中,输入的视觉特征通过第一层 MLP 进行线性变换,经过激活函数后进入第二层,映射到与语言嵌入相同的维度,其映射关系如式(2)所示:

$$\mathbf{H}_V = \mathbf{W}_2(\sigma(\mathbf{W}_1 \mathbf{Z}_V + b_1)) + b_2 \quad (2)$$

式中: $\mathbf{H}_V$ 是图像特征; $\mathbf{W}_1$ 和 $\mathbf{W}_2$ 分别是第一、第二层的权重矩阵; b 是偏置项; $\sigma(\cdot)$ 是非线性激活函数。

2.1.3 大语言模型

预训练大语言模型(LLM)负责对融合后的视觉与语言特征进行深度理解和智能响应生成。本文延续了 LLaVA 采用的 Vicuna 模型^[17]作为核心的 LLM 处理组件,这是一种基于 Transformer 架构的自回归语言模型,其经过大量文本数据的预训练,具备强大的自然语言理解、逻辑推理和连贯生成能力。在与视觉特征结合时,Vicuna 能够利用其语言理解能力,在充分考虑图像信息的基础上,生成准确且富有信息的回答,确保对复杂街景场景的精确描述和

推理。

在具体运行过程中,经过 MLP 连接器处理后的视觉特征和文本指令一起输入 LLM。通过 LLM 的多层 Transformer 编码器对这些融合特征进行处理,逐步生成自然语言响应。Vicuna 运用自回归机制,依据上下文信息逐词构建答案,有力保障了响应的连贯性与准确性。模型融合特征并生成对应描述过程可表示为式(3):

$$\mathbf{X}_a = f_\phi(\mathbf{H}_V, \mathbf{H}_q) \quad (3)$$

式中: $\mathbf{H}_q$ 是提示指令的文本特征; $f_\phi(\cdot)$ 是 LLM 的映射函数; $\mathbf{X}_a$ 是最终生成的描述。

2.2 街景视觉-指令跟随数据集构建

指令跟随(Instruction Following)是大模型评估中模型能够准确理解和执行用户提供的指令的能力。在本研究中,为了增强模型对街景场景分析相关指令的理解和执行能力,构建了基于街景场景的视觉-指令跟随数据集,如图 2 所示。该数据集包括街景图像及其地理信息、提示指令库,以及利用 GPT-4 辅助生成相应的标准描述后的视觉-指令跟随数据。

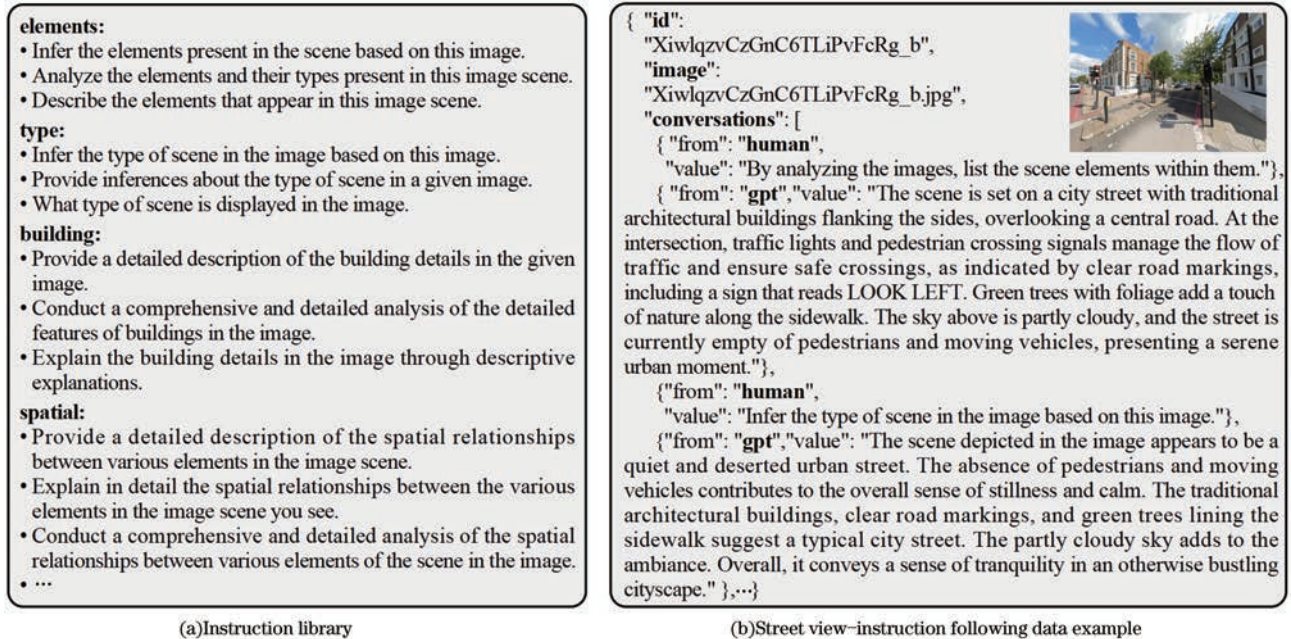


图 2 街景视觉-指令跟随数据集

Fig.2 Street view-instruction following dataset

2.2.1 提示指令设计

为确保模型尽可能全面地描述场景,本文设计了 4 个提示框架,每个框架专注于从提取角度分析街景场景:“elements”框架用于识别场景中的对象;“type”框架用于推断场景类型;“spatial”框架用于描述对象之间的空间关系;“building”框架聚焦于建筑物的细节特点。这些框架覆盖了场景中多个层次

的要素,如具体对象、场景类型、建筑物细节和空间关系,旨在增强模型对场景上下文的全面理解。图 2(a)展示了构建的部分提示指令库,每个框架下都包含多种表述形式相似但措辞不同的指令,并使用“infer”“Describe”“Analyze”等引导词。这些指令通过多样化的表达方式,引导模型深入挖掘并准确提取场景信息。

2.2.2 基于 GPT 的标准答案生成

为了构建高质量的训练数据集,本文采用了一种半自动化的方法来生成标准场景描述。这一过程利用 GPT-4 进行模拟问答,对于训练集中的每幅图像 $\mathbf{X}_v$,询问 2~4 个提示指令 $\mathbf{X}_q$,模型根据图像内容生成相应的答案 $\mathbf{X}_c$,从而构建多轮对话数据。对话数据被组织成一个序列,每个对话轮次 t 包括一个提示指令 $\mathbf{X}_{q,t}$ 及其对应的标准回答 $\mathbf{X}_{c,t}$,格式为:

human: $\mathbf{X}_{q,1}\mathbf{X}_v<\text{STOP}>\text{Assistant: } \mathbf{X}_{c,1}<\text{STOP}>$
 human: $\mathbf{X}_{q,2}<\text{STOP}>\text{Assistant: } \mathbf{X}_{c,2}<\text{STOP}>$
 $\vdots$
 human: $\mathbf{X}_{q,T}<\text{STOP}>\text{Assistant: } \mathbf{X}_{c,T}<\text{STOP}>$

式中: T 是对话的总轮数; $<\text{STOP}>$ 表示当前回答的结束。通过进一步的人工检查与改进,共获得 1 054 对高质量的指令跟随数据,具体的示例如图 2(b)所示。

2.3 模型微调

本文模型在微调前进行了预训练,该过程参照了 LLaVA 模型的预训练流程与数据集^[14],仅更新 MLP 映射层的参数,实现图像特征与 LLM 词嵌入的对齐。

2.3.1 LoRA 微调策略

在对 GeoLLaVA 进行微调时,本研究面临诸多挑战。一方面,LLM 因参数量巨大,全参数微调对计算资源和存储空间要求极高;另一方面,为确保训练数据质量,人工辅助制作数据集致使数据规模受限,这可能引发模型训练时的过拟合问题。

为解决以上问题,本文引入 LoRA,相较于其他微调方法,LoRA 通过引入低秩可训练层,在保持预训练模型大部分权重固定的同时,有效增强了模型的微调能力,从而提高了模型的性能和泛化能力^[40]。具体地,LoRA 冻结预训练模型的主要权重后,将可训练的低秩分解矩阵嵌入每个 Transformer 模块中,通过降维再升维的方式调整参数。该方法显著减轻了梯度计算负担,降低了图形处理器(GPU)内存需求,并有效降低过拟合风险,使得模型在有限资源和数据条件下高效微调,且保持较强的泛化能力。LoRA 微调过程如图 3 所示。

LoRA 使用两个与原模型并行的低秩矩阵:降维矩阵 $\mathbf{A}$ 和升维矩阵 $\mathbf{B}$,替代增量权重矩阵 $\Delta\mathbf{W}$,模型训练不直接微调 LLM 的预训练参数 $\mathbf{W}_0$,只更新 $\mathbf{A}$ 和 $\mathbf{B}$ 的参数。在前向过程中, $\mathbf{W}_0$ 和 $\Delta\mathbf{W}$ 都会与相同的输入 x 相乘,最后相加(类似残差连接)得

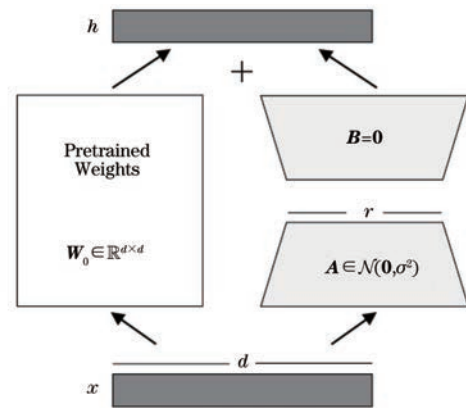


图 3 LoRA 方法训练过程

Fig.3 LoRA method training process

出,如式(4)所示:

$$h = \mathbf{W}_0 x + \Delta\mathbf{W}x \approx \mathbf{W}_0 x + \mathbf{B}\mathbf{A}x \quad (4)$$

式中: $\mathbf{W}, \Delta\mathbf{W} \in \mathbb{R}^{d \times d}$, $\mathbf{A} \in \mathbb{R}^{r \times d}$ 初始化为正态分布; $\mathbf{B} \in \mathbb{R}^{d \times r}$ 初始化为 $\mathbf{0}$; $\Delta\mathbf{W}$ 的秩 $r \ll d$ 。

LoRA 的微调质量与全模型微调相当,但仅需更新万分之一的参数,显著减少了训练时间和计算资源消耗,尤其适用于大型模型。此外,对于不同的下游任务,仅需对预训练 LLM 中的 $\mathbf{A}$ 和 $\mathbf{B}$ 进行重新训练,有效提高了模型适应新任务的速度和灵活性。

2.3.2 视觉指令微调

为了使模型能够适用于本研究中的街景数据并输出理想的描述结果,本文在视觉-指令跟随数据上对模型进行端到端微调。模型的可训练参数为 $\theta = \{\mathbf{W}, \phi\}$,即冻结视觉编码器的权重,专注于训练 MLP 层与 LLM 的参数。对于每张图像 $\mathbf{X}_v$,生成多轮对话数据 $(\mathbf{X}_{q,1}\mathbf{X}_{a,1}, \dots, \mathbf{X}_{q,T}\mathbf{X}_{a,T})$,其中 T 为总对话轮数。将多轮对话组织为一个序列,将所有的回答视为模型的回答结果。指令 $\mathbf{X}_{\text{instruct},t}$ 在第 t 轮中根据以下方式生成:当 $t=1$ 时,随机选择 $[\mathbf{X}_{q,1}, \mathbf{X}_v]$ 或 $[\mathbf{X}_v, \mathbf{X}_{q,1}]$ 作为初始提示;当 $t>1$ 时,使用相应的 $\mathbf{X}_{q,t}$ 作为提示。在训练过程中,针对每个序列,根据长度 L 计算目标回答 $\mathbf{X}_a$ 的概率,如式(5)所示:

$$p(\mathbf{X}_a | \mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^L p_{\theta}(x_i | \mathbf{X}_v, \mathbf{X}_{\text{instruct}, <i>, \mathbf{X}_{a, <i>}) \quad (5)$$

式中: θ 为可训练参数; $\mathbf{X}_{\text{instruct}, <i>$ 和 $\mathbf{X}_{a, <i>$ 分别为指令和答案在第 i 个预测标记前的所有轮次的序列。在条件概率计算中,显式地引入图像 $\mathbf{X}_v$,以强调每个回答都是基于图像内容生成的。对于每个影像,模型接收多条提示指令,并基于这些指令和图像内容进行推理,生成相应的描述和结果标签。

在训练过程中,模型生成的结果标签与由 GPT-4 提供的标准答案标签进行比对。为了计算损失值,将生成的结果标签与标准答案标签分别扁平化为一维张量,通过交叉熵损失函数^[41]计算两者之间的差异,以评估模型在当前参数设置下的性能。基于所得到的损失值,模型通过反向传播计算梯度,进而指导模型参数的优化调整,旨在逐步减少损失值,从而优化模型性能。该过程在多个训练周期中反复进行,直至模型的性能达到预期标准。

2.4 场景典型要素提取

在模型推理过程中,从指令库的各个框架中随机选取一条指令,将其与图像进行配对,从而形成“图片-指令对”,随后将这些配对数据加载至模型中,模型依据所提供的提示对图像展开描述。

在推理过程中,GeoLLaVA 生成的结果与输入的原始图像、指令框架及地理位置信息精准关联。基于以上推理结果构建场景要素的初始数据库,该数据库为后续研究提供了至关重要的数据支撑以及分析基础。为了从推理得到的场景描述中准确提取出特定的场景要素,例如停车场(parking lot)、路口(intersection)、交通信号灯(traffic lights)等,本文采取了一系列文本处理和分析的操作,以确保信息的完整性和准确性。

首先,对描述信息进行预处理,包括词干化、去除停用词以及分词。

1)词干化:将词汇还原到基本形式的过程,如将“泊车区”“停车坪”等变换为“停车场”,以提高处理的准确性。

2)去除停用词:删除在信息提取中不具实质意义的常见词汇,如“的”“和”“是”等,以减少文本中的噪声。

3)分词:将连续的文本分为独立词汇或短语。

然后,进行关键信息提取,包括关键词和短语识别、上下文分析以及标签生成。

1)关键词和短语识别:如为了提取交通信号灯要素,搜索包含“traffic lights”“green light”等词汇,提取停车场要素时检索“parking space”等描述性短语。

2)上下文分析:考虑描述信息的整体上下文,以识别潜在的地理信息要素,例如,即使文本中没有直接提到“停车场”,但若有描述车辆停在空地上的情境,也可推测存在停车场。

3)标签生成:为包含特定地理信息要素的数据打上标签,并将其对应的经纬度信息输出。

3 实验与结果分析

3.1 实验设置

3.1.1 实验数据

本研究使用的街景图片来自 GSV 数据集,均位于英国伦敦城市区域。如图 4 所示,本文将原始全景街景转换为 4 个水平视角的街景图像,避免全景图像的视角变形和压缩导致模型理解的偏差,同时保证了每个地理位置街景场景的完整,避免典型场景要素的遗漏。



图 4 街景全景数据的预处理结果

Fig. 4 Pre-processing results for street view panorama data

模型的预训练使用的是 LLaVA Visual Instruct CC3M Pretrain 595K 数据集,包含来自图像-字幕对的多模态合成对话。对于模型的微调,本文构建了 1 054 对街景视觉-指令跟随数据。为了定量评估模型的街景要素提取能力,本文通过人工判断的方式,为 1 000 张测试图像标注街景典型要素标签,包括交通信号灯、路口、停车场,每一个要素标签表示图像场景存在对应的视觉实体。

此外,本研究还进行了区域应用实验,使用了一片面积约 4.5 km²、包含 20 692 张街景图像的完整区域数据,并以对应范围的 OSM 数据作为参考。

3.1.2 评价指标

为定量评估模型提取街景图像典型场景要素的能力,本文采用的评估指标^[27,42]包括召回率(R)、精确率(P)和 F1 值(F₁)。在实验中,模型对测试图像推理并输出场景描述,然后将推理结果中提取的要素和真实标注对比,统计出相应的真阳性(TP)、假阳性(FP)、真阴性(TN)和假阴性(FN)的数量 N_{TP} 、 N_{FP} 、 N_{TN} 、 N_{FN} ,进而计算各指标,以此把握模型在该任务中的性能表现,以及准确性、全面性的平衡情况。

精确率衡量模型在判定为正例的样本中,正确识别正例的比例,高精确率表明模型具有较低的误判率和较高的判定可靠性。精确率计算公式如式(6)所示:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \quad (6)$$

召回率衡量模型正确识别的正例占所有实际正例的比例,反映模型在街景图像中识别要素的能力。较高的召回率意味着模型遗漏实际要素的情况较少,能够更全面地覆盖需提取的要素。召回率计算公式如式(7)所示:

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad (7)$$

F1 值是精确率和召回率的调和平均数,综合考虑了模型的准确性与全面性。F1 值计算公式如式(8)所示:

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (8)$$

3.1.3 实现细节

模型的预训练与微调训练是在 1 张 A800 GPU (80 GB 显存)上完成的。预训练过程采用余弦退火学习率调度策略,学习率为 2×10^{-3} ,训练周期为 1,批量为 16,梯度累积步数为 1,全局批量大小为 256。微调过程采用余弦退火学习率调度策略,预热比率设定为 0.03,学习率为 2×10^{-5} ,训练周期为 1,批量大小为 16,梯度累积步数为 1,全局批量大小为 128。最大序列长度为 2 048,超参数设置遵循 Vicuna 的配置^[9]。

在推理阶段,实验设定了 4 个数据加载器工作线程,使用了 1 张 A800 和 3 张 A10 GPU 进行并行处理,每个 GPU 的批量大小设定为 200。

3.2 对比实验

3.2.1 模型性能对比

本文选取街景要素提取领域具有代表性的多类

模型开展定量对比分析,涉及经典语义分割、目标检测方法以及前沿多模态模型方法,分别测试各方法在街景图像上提取交通信号灯、交叉路口和停车场要素的性能,结果如表 1 所示(最优指标值加粗标注,下同)。

具体基线模型如下:

1)语义分割方法:在 Cityscapes 数据集上训练的 PSPNet 和 DeepLabV3+ 模型,主干网络选用 ResNet101,输出结果为像素级的分类,本文将像素数量超过 50 的“traffic lights”类别作为有效识别要素。

2)目标检测方法:在 COCO(Microsoft Common Objects in Context)数据集上训练的 YOLOv8^[43] 和 YOLO11^[44] 模型,输出结果为目标框、对象类别和置信度,本文设定置信度大于 0.25^[43] 的“traffic lights”实例作为有效识别要素。

3)多模态方法:包括预训练 CLIP 模型(ViT-L/14 和 ViT-L/14@336px 版本),通过计算图像和文本的相似度,将图像与预设的标签匹配:labels = ["There are traffic lights", "There is a parking lot", "There is an intersection"],以相似度分数高于 0.5^[45] 的标签作为有效识别要素;还包括大型多模态模型 InternVL2^[46]、LLaMA3^[47]、Qwen2^[48]、Pixtral^[49] 和 Gemini-1.5^[50]。

为确保对比实验的合理性,本研究选取与 GeoLLaVA 参数规模相近的百亿参数模型版本,并借助 Chatbot Arena 平台^[51]对这些基线模型予以测试,其推理流程与本研究类似,即输入图片与提示以获得答案后提取要素。

表 1 本文方法与其他方法在街景要素提取任务中的性能对比

Table 1 Comparison of our method with other methods in scene element extraction task

Methods	traffic lights			intersection			parking lot		
	P	R	F ₁	P	R	F ₁	P	R	F ₁
PSPNet (ResNet101)	0.844	0.876	0.860	—	—	—	—	—	—
DeepLabV3 (ResNet101)	0.881	0.752	0.811	—	—	—	—	—	—
YOLOv8	0.959	0.679	0.795	—	—	—	—	—	—
YOLO11x	0.971	0.679	0.799	—	—	—	—	—	—
CLIP (ViT-L/14)	0.790	0.133	0.228	0.552	0.753	0.637	0.278	0.612	0.382
CLIP (ViT-L/14@336px)	0.849	0.475	0.609	0.641	0.661	0.651	0.235	0.680	0.349
Qwen2-VL-7B-Instruct	0.882	0.455	0.600	0.667	0.098	0.170	0.750	0.360	0.487
Gemini-1.5-Flash-8B	0.857	0.546	0.667	0.711	0.659	0.684	0.667	0.720	0.692
llama-3.2-vision-11b-instruct	0.842	0.485	0.615	0.571	0.195	0.291	0.769	0.400	0.526
Pixtral-12B	0.895	0.515	0.654	0.546	0.146	0.231	0.438	0.280	0.342
InternVL2-26B	0.957	0.667	0.786	0.533	0.195	0.286	0.813	0.520	0.634
GeoLLaVA	0.905	0.973	0.938	0.860	0.825	0.842	0.889	0.777	0.829

在交通信号灯提取任务中,YOLOv8 和 YOLO11x 的精确率较高,分别达到 0.959 和 0.971,

但召回率明显较低,存在较多漏检情况。PSPNet 和 DeepLabV3+ 的表现较为稳定,但整体表现不及

GeoLLaVA。CLIP 系列多模态模型和大型多模态模型在交通信号灯提取时,精确率和召回率表现不一,多存在精确率尚可但召回率较低的问题,整体性能一般。其中 InternVL2 模型获得了 0.957 的精确率,这可能得益于该模型更大规模的参数量级。而本文提出的 GeoLLaVA 模型在交通信号灯提取上实现了精确率(0.905)和召回率(0.973)的良好平衡,F1 值达到 0.938,显著优于其他基线模型。

由于训练数据的局限性,上述语义分割方法和目标检测方法在本文设定任务中仅能完成交通信号灯的提取。交叉路口和停车场等要素并不具备统一的视觉特征,而是由多个元素(如道路交会、停车位、车道划分等)组成,这要求模型具备理解这些要素间空间和功能关系的能力。然而,现有的图像分割和目标检测方法并未针对这些复杂场景中的要素提取进行专门的研究,因此无法有效处理这类任务,导致在这些要素的提取上缺乏相关的研究进展。

在交叉路口和停车场提取任务中,GeoLLaVA 均展现出较强竞争力,各项指标得分最高。这表明 GeoLLaVA 能更有效地理解和提取街景图像中的复杂场景要素,优于其他传统视觉方法和多模态大模型。

具体来说,在交叉路口要素提取任务中,基线模型整体表现较差,如 Qwen2 的召回率低至 0.098,

即便是表现较好的 Gemini,其 F1 值也仅有 0.684,远未达到理想水平。同样,在停车场要素提取任务中,CLIP 系列模型平均精确率仅 0.257,参数量最大的 InternVL2 精确率达到 0.813,但召回率只有 0.520,整体性能较差。以上结果显示,经大规模数据预训练的多模态模型在复杂典型场景要素提取任务中难以满足实际应用需求。CLIP 模型虽然在整体表现上与其他大型模型没有显著差异,但其依赖于图像和文本之间的相似度匹配机制,预设标签和相似度阈值的选择直接影响了提取效果,这种相似度匹配机制使其无法适应多样化的要素表达。

尽管上述大型多模态模型具有丰富的预训练知识和较强的上下文理解能力,但其训练数据并没有特别关注或强化对复杂场景要素的提取能力。因此,即使这些模型在广泛的场景中能够发挥一定的优势,但在面对复杂多样的要素时,缺乏足够的训练和优化。本文所提出的 GeoLLaVA 模型注重在训练数据中针对性地加强对这些复杂场景要素的学习,确保模型能够在这些复杂任务中充分理解并准确提取关键要素,因此,GeoLLaVA 在这些复杂要素提取任务中的表现明显优于其他模型。

3.2.2 微调前后对比分析

模型微调前后对街景图像的描述能力如表 2 所示,其中,“…”部分是被删去部分的描述,在本文的

表 2 模型微调前后描述信息对比

Table 2 Comparison of descriptive information before and after model fine-tuning

SVI Data	Model Output Before Fine-tuning	Model Output After Fine-tuning
	【elements】 The image depicts a busy street with several cars parked along the side of the road. ... In addition to the cars, there are two trucks parked along the street. ... There are also three people in the scene. The presence of pedestrians and parked vehicles suggests that this is a bustling urban area with active traffic and parking.	【elements】 The image shows a residential street with many parked cars, several trees, and a street lamp on both sides of the road. On the right side of the road is a residential area where brown houses can be seen. There is a parking lot on the left side of the road, where at least four cars are parked, and there is also a lush garden in the background. ...The image is a photo taken horizontally from a ground perspective...The image is in color, and the colors are bright and vivid. The image is in focus, and the details are clear and sharp. ...
	【type】 The image displays a scene of a large, fancy house with a white fence and a black gate. ...	【type】 The scene in the image appears to be from an affluent neighborhood. ... It seems to be a peaceful and upscale urban or suburban area.
	【building】 The image features a city street with a row of white buildings on one side. The buildings are tall and have a uniform appearance, likely part of a residential or commercial complex. ... There are several people walking on the sidewalk, going about their daily activities. ...	【building】 The image shows a street in London, with a number of buildings and a few cars parked on the side of the road. The houses are all white and have a number of windows on each floor, with a few balconies on the top floor. The roofs are all flat and have a number of chimneys. ...There are a few trees on the side of the road, and the pavement is made of concrete. ...There is a sidewalk in front of the house, with many street lights and a red mailbox on it. ...
	【spatial】 The image shows a residential street in a city, with several buildings and a few cars parked on the side of the road... The street is paved with asphalt and has a few trees and bushes along the sides. ...	【spatial】 The image depicts a quiet residential street with several cars parked along the side of the road. ...The street is lined with brick buildings, and there are trees and a bench nearby, adding to the overall ambiance of the scene. The street is situated between two brick buildings, with one building on the left side and the other on the right side. The parked cars are positioned in a way that they do not obstruct the view of the buildings, and they are not too close to each other, allowing for a comfortable parking arrangement.

关注点之外,例如“The sky is overcast and there are a few clouds in the background.”描述天空和云,与本文研究无关。

在场景元素描述方面,微调前模型主要关注街道上的汽车和行人,强调交通和区域的繁忙程度。微调后模型不仅能识别出汽车、道路、树木、公园停车场和路灯,还能提供更多视觉细节,如拍摄角度和色彩饱和度。在场景类型分析上,微调前模型并没有进行关于场景类型的推理,而是停留在建筑物层面,微调后则是开始推测场景可能是位于郊区的高档小区。在建筑物描述上,微调前模型描述较为概括,主要提及建筑物颜色,并粗略推测其功能类型。微调后模型则提供了涵盖窗户、阳台、屋顶、烟囱等更为细致的描述。此外,它还涉及街道环境中的元素,如汽车、树木、人行道、路灯和邮筒,并结合视角和天气条件对场景进行了全面的描述。在空间结构描述上,微调前模型的描述较为简单,仅说明街道、建筑物及车辆的基本布局。微调后模型则精确地描述了

街道两旁建筑物的排列方式及其空间关系,如树木与建筑物之间的布局以及街道上物体的相对位置。

GeoLLaVA 模型在微调后,不仅能更全面地识别并描述图像中的自然和人造元素,在要素细节以及空间关系的描述上展现出了更强的丰富性和准确性,还对不同指令有了更深刻的理解,能根据指令提供不同视角的场景分析。

3.3 消融实验

3.3.1 模型改进结构消融实验

为了验证提出的模型改进结构在街景图像典型场景要素提取任务中的有效性,本文进行了消融实验,测试了不同结构模型在经过微调后的要素提取结果。

从表 3 中可以看出,综合了 ViT-L/14@336px 的视觉编码器和 MLP 映射层改进结构的 GeoLLaVA 模型在各要素提取任务上均达到了更优的性能表现,进一步证明了模型改进结构的有效性和优势,为街景图像典型场景要素提取任务提供了更高效、精准的解决方案。

表 3 模型改进结构消融实验评价指标

Table 3 Evaluation metrics in ablation experiment of improved model structures

Models	traffic lights			intersection			parking lot		
	P	R	F ₁	P	R	F ₁	P	R	F ₁
LLaVA	0.857	0.851	0.854	0.823	0.763	0.792	0.487	0.728	0.584
LLaVA+MLP	0.860	0.874	0.867	0.828	0.788	0.807	0.517	0.748	0.611
LLaVA+ViT-L/14@336px	0.878	0.942	0.909	0.856	0.801	0.828	0.828	0.748	0.786
GeoLLaVA	0.905	0.973	0.938	0.860	0.825	0.842	0.889	0.777	0.829

具体地,MLP 映射层相比单层线性映射,能够更有效地捕捉复杂的映射关系,增强了模型对视觉与语言信息的融合能力,这在各要素提取任务的指标提升上也有体现。相较于基线模型中输入分辨率为 224×224 像素的视觉编码器,ViT-L/14@336px 将输入图像分辨率提升至 336×336 像素,能够提供更丰富的视觉细节,帮助 LLM 更精准地感知高清街景图像的内容,例如在停车场提取任务中,相较于基线模型 LLaVA,精确率提升了 0.341,F1 值提升了 0.202。

3.3.2 LoRA 消融实验

为了进一步验证 LoRA 在降低计算资源消耗方面的有效性,本研究开展了相关的消融实验。通

过将 LoRA 微调策略与全参数微调方法进行对比,测试 LoRA 对模型性能以及计算资源消耗的影响。实验在 2 张 A800 GPU(80 GB 显存)上进行,为了在降低每个训练步骤计算负担的同时维持训练的稳定性,对训练参数进行了调整,训练批量大小由 16 降低至 1,同时将梯度累积步数从 1 增加到 16。

实验结果如表 4 所示。引入 LoRA 后,模型训练时间显著缩短,仅为全参数微调训练时间的 32%,所需显存量大幅降低,约为全参数微调时的 25%。在街景要素提取任务中,LoRA 微调后的模型性能与全参数微调的模型表现相当。这一结果充分验证了 LoRA 在保证性能的前提下,能够显著提高训练效率和资源利用率。

表 4 LoRA 消融实验评价指标

Table 4 Evaluation metrics in ablation experiment of LoRA

Fine-tuning Methods	traffic lights			intersection			parking lot			Time/h	GPU Memory Usage/GB
	P	R	F ₁	P	R	F ₁	P	R	F ₁		
Full-parameter	0.907	0.973	0.939	0.861	0.817	0.839	0.899	0.777	0.833	4.069	122.499
LoRA	0.905	0.971	0.937	0.862	0.820	0.838	0.870	0.777	0.821	1.287	30.932

3.4 区域应用实验

3.4.1 模型推理结果与可视化

本研究将特定区域的街景图像输入 GeoLLaVA 大模型进行批量推理,模型根据输入图

像生成全面的场景描述。模型生成的描述如图 5 所示,涵盖了“element”“type”“building”“spatial”4 个维度的信息,为每一张街景图像提供了详尽的场景理解。

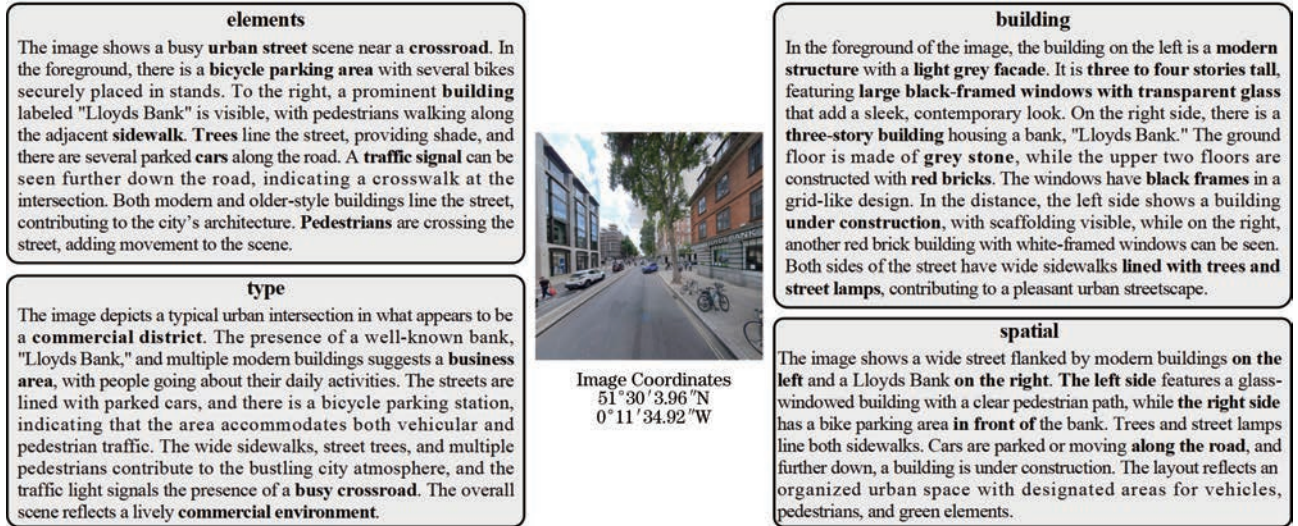


图 5 GeoLLaVA 生成的街景影像描述信息

Fig.5 Street view image description information generated by GeoLLaVA

通过关键词提取典型场景要素(如交通信号灯、路口、停车场等),本文进一步对提取的要素进行了精确标注,结合每个要素的经纬度信息,绘制了地图

上的可视化图示。图 6 展示了不同类型的街景要素分布与具体案例,使用不同符号区分这些要素,清晰地展示了各类要素在特定区域的分布情况。



图 6 GeoLLaVA 提取的要素分布及具体案例

Fig.6 Distribution of elements extracted by GeoLLaVA and description of a specific location

本文将提取的要素与地理坐标结合,不仅清晰展示了这些地理要素的具体位置,还揭示了它们在城市空间中的分布模式与特点。例如,停车场通常集中于商业区和居民区等人流、车流密集的区域,而路口和交通信号灯多分布于道路网络的节点处,以合理地引导和控制交通流量。这一过程不仅展示了

GeoLLaVA 在区域实际应用中的有效性,还为后续的城市规划、交通管理、地理定位等领域提供了有力的数据支持。

3.4.2 模型提取与 OSM 数据对比分析

OSM 的数据主要来源于 GPS 轨迹、航拍影像以及社区成员的贡献,提供了精确的道路网络信息

和直观的地图呈现。本研究通过对 GeoLLaVA 提取的地理信息与 OSM 平台的标注信息进行定性对比分析,旨在揭示两者之间的相似性和差异,验证模型的准确性,并识别 OSM 可能存在的信息遗漏或错误。图 7 展示了交通信号灯、路口和停车场的分布与 OSM 标记的对比分析。如图 7(a)所示,两者在主干道上的信号灯分布基本一致,但在建筑密集区,GeoLLaVA 的识别更为准确。这可能是由于 OSM 忽略了部分住宅区信号灯或未及时更新,而 GeoLLaVA 能够识别并标记这些设施。如图 7(b)所示,GeoLLaVA 提取的路口位置与 OSM 几乎一致,可能是因为路口处于开阔地带,易于捕捉。然

而,部分 OSM 标识的路口未被模型识别,可能是由于街景影像不连续、视角盲区或信息模糊处理影响了模型的推理。然而,即使从 OSM 能够分辨出路口的分布,但其统计过程需要依靠人工辨认,基于模型自动化的提取方法则更具优势。如图 7(c)所示,GeoLLaVA 提取的停车场数量明显多于 OSM 标记。未标记的停车场大多位于树木周围或道路两侧的临时停车区。街景图像清晰显示了这些被遮挡的停车位,模型因此将其识别为停车场,而 OSM 地图则往往将这些区域标记为普通道路,而非基于实际使用情况的动态识别。此外,住宅区的车库或停车位虽未在 OSM 中标记,但被 GeoLLaVA 识别出。

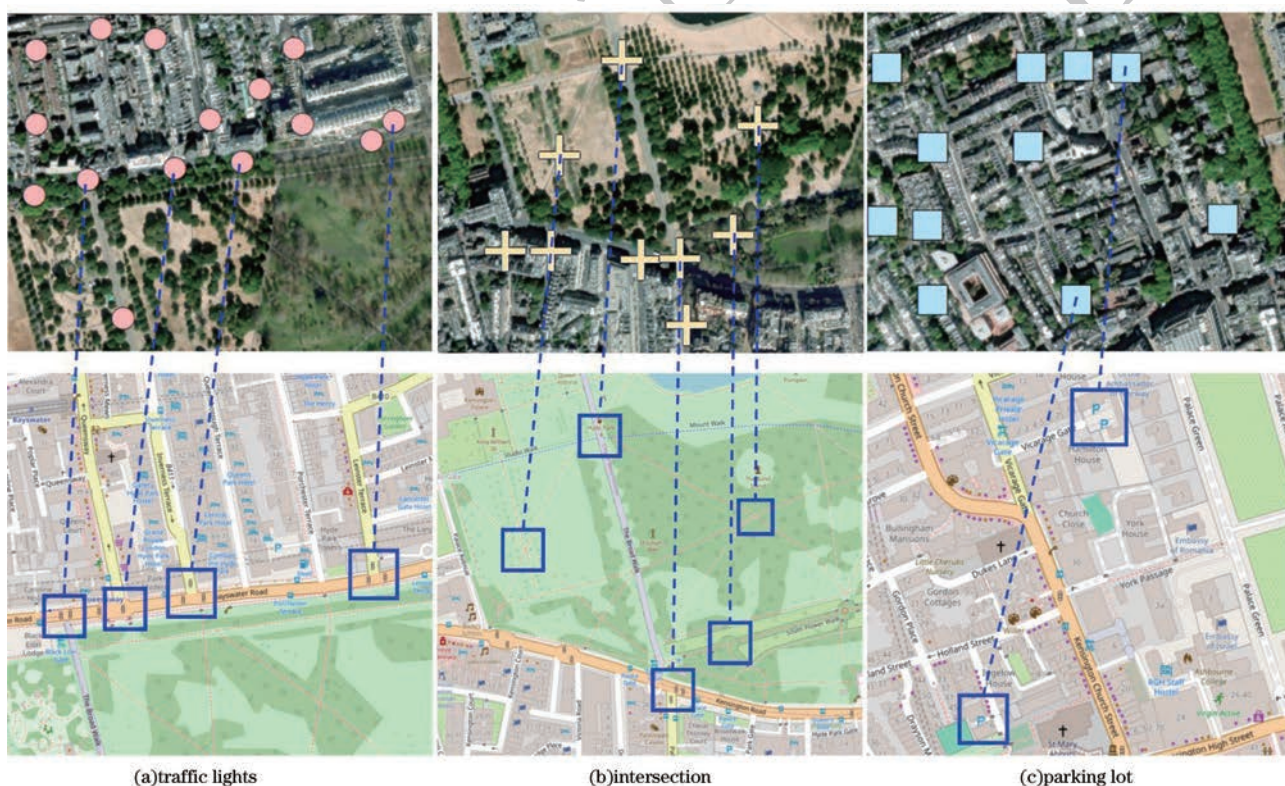


图 7 典型场景要素可视化分布与 OSM 结果对比

Fig.7 Visualization of scene elements distribution and comparison with OSM results

模型识别结果与 OSM 数据的差异可能受多种因素影响,包括算法的差异、数据源的更新频率以及地图制作者的主观判断。为了进一步提升地理信息的准确性,未来可以结合多种数据源和识别技术,不断优化模型,以获得更全面的地理要素分布信息。

4 结束语

本文提出了一种基于大型多模态模型的街景图像典型场景要素提取方法,通过构建 GeoLLaVA 模型并利用 LoRA 微调,实现了街景要素的精确提取。本文专注于提升模型在复杂街景场景中对关键地理要素的提取能力,尤其是在交通信号灯、交叉路

口和停车场等要素的识别任务中取得了显著的性能提升。同时,本文构建了街景视觉-指令跟随数据集,致力于通过多模态方法实现街景场景地理信息的全面描述。实验结果表明,该方法在与多种模型和方法的对比中表现优异,且在区域应用中能有效提取要素并揭示其分布模式,同时也揭示了现有 OSM 数据在要素提取方面的不足之处。尽管如此,本研究在进一步提升要素提取性能以及应对更多样化场景要素的识别任务上仍有潜力和进步空间。在未来的研究中,计划进一步改进多模态融合机制,优化模型中不同模块之间的交互方式,在视觉编码器、映射层和语言模型之间建立更紧密的反馈机制,以

提高各模块之间的信息流动和更新效率,从而提升模型对街景要素的整体理解能力。此外,还计划扩展视觉-指令跟随数据集的规模和多样性,增加更丰富的任务指令,进一步增强模型在复杂地理信息提取任务中的鲁棒性和泛化能力。

参考文献

- [1] HJELM R D, FEDOROV A, LAVOIE-MARCHILDON S, et al. Learning deep representations by mutual information estimation and maximization [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/1808.06670>.
- [2] 张丽英, 裴韬, 陈宜金, 等. 基于街景图像的城市环境评价研究综述[J]. 地球信息科学学报, 2019, 21(1): 46-58.
ZHANG L Y, PEI T, CHEN Y J, et al. A review of urban environmental assessment based on street view images[J]. Journal of Geo-Information Science, 2019, 21(1): 46-58. (in Chinese)
- [3] DOERSCH C, SINGH S, GUPTA A, et al. What makes Paris look like Paris? [J]. Communications of the ACM, 2015, 58(12): 103-110.
- [4] NGUYEN Q C, SAJJADI M, MCCULLOUGH M, et al. Neighbourhood looking glass: 360° automated characterisation of the built environment for neighbourhood effects research[J]. Journal of Epidemiology and Community Health, 2018, 72(3): 260-266.
- [5] ZHANG F, ZHANG D, LIU Y, et al. Representing place locales using scene elements[J]. Computers, Environment and Urban Systems, 2018, 71: 153-164.
- [6] DEWI C, CHEN R C, ZHUANG Y C, et al. Image enhancement method utilizing YOLO models to recognize road markings at night[J]. IEEE Access, 2024, 12: 131065-131081.
- [7] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE Press, 2017: 6230-6239.
- [8] LIANG X C, ZHAO T H, BILJECKI F. Revealing spatio-temporal evolution of urban visual environments with street view imagery[J]. Landscape and Urban Planning, 2023, 237: 104802.
- [9] XU S X, ZHANG C H, FAN L B, et al. AddressCLIP: empowering vision-language models for city-wide image address localization[C]//Proceedings of ECCV 2024. Berlin, Germany: Springer, 2025: 76-92.
- [10] NGIAM J, KHOSLA A, KIM M, et al. Multimodal deep learning[C]//Proceedings of the International Conference on Machine Learning (ICML). [S. l.]: PMLR, 2011: 689-696.
- [11] HE W T, MA H J, LI S H, et al. Using augmented small multimodal models to guide large language models for multimodal relation extraction[J]. Applied Sciences, 2023, 13(22): 12208.
- [12] OUYANG T J, ZHANG X, HAN Z Y, et al. Health CLIP: depression rate prediction using health related features in satellite and street view images[C]//Proceedings of the ACM Web Conference 2024. New York, USA: ACM Press, 2024: 1142-1145.
- [13] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical report[EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2303.08774>.
- [14] LEE Y J, LI C Y, LIU H T, et al. Visual instruction tuning [J]. Advances in Neural Information Processing Systems, 2023, 23: 34892-34916.
- [15] ZHANG Y, ZHANG F, CHEN N C. Migratable urban street scene sensing method based on vision language pre-trained model [J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 113: 102989.
- [16] JI Y H, GAO S. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2307.03678>.
- [17] CHIANG W L, LI Z, LIN Z, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90% * ChatGPT quality [EB/OL]. [2024-12-19]. <https://lmsys.org/blog/2023-03-30-vicuna>.
- [18] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of International Conference on Machine Learning. [S. l.]: PMLR, 2021: 8748-8763.
- [19] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2106.09685>.
- [20] BAI Y, ZHAO Y, SHAO Y J, et al. Deep learning in different remote sensing image categories and applications: status and prospects [J]. International Journal of Remote Sensing, 2022, 43(5): 1800-1847.
- [21] 徐永智. 基于街景影像的建筑物底部轮廓提取[D]. 北京: 北京建筑大学, 2017.
XU Y Z. Extracting building footprints from digital measurable images[D]. Beijing: Beijing University of Civil Engineering and Architecture, 2017. (in Chinese)
- [22] LU J S, YANG J W, BATRA D, et al. Neural baby talk[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Press, 2018: 7219-7228.
- [23] CAMPBELL A, BOTH A, SUN Q. Detecting and mapping traffic signs from Google Street View images using deep learning and GIS[J]. Computers, Environment and Urban Systems, 2019, 77: 101350.
- [24] QIU W S, ZHANG Z Y, LIU X, et al. Subjective or objective measures of street environment, which are more effective in explaining housing prices? [J]. Landscape and Urban Planning, 2022, 221: 104358.
- [25] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. [2024-12-19]. <https://arxiv.org/abs/1706.05587>.
- [26] SHIHAB I F, BHAGAT S R, SHARMA A. Precise and robust sidewalk detection: leveraging ensemble learning to surpass LLM limitations in urban environments[EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2405.14876>.
- [27] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE Press, 2017: 6517-6525.
- [28] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE Press, 2016: 3213-3223.
- [29] ZHOU B L, ZHAO H, PUIG X, et al. Semantic understanding of scenes through the ADE20K dataset [J]. International Journal of Computer Vision, 2019, 127(3): 302-321.
- [30] ZHANG Y, LIU P Y, BILJECKI F. Knowledge and topology: a two layer spatially dependent graph neural networks to identify urban functions with time-series street view image [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 198: 153-168.
- [31] WU M L, HUANG Q Y, GAO S, et al. Mixed land use

- measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning [J]. *International Journal of Applied Earth Observation and Geoinformation*, 2023, 125: 103591.
- [32] ZHAO Y H, ZHONG E H, YUAN C Y, et al. TG-LMM: enhancing medical image segmentation accuracy through text-guided large multi-modal model [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2409.03412>.
- [33] PICARD C, EDWARDS K M, DORIS A C, et al. From concept to manufacturing: evaluating vision-language models for engineering design [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2311.12668>.
- [34] YANG Y F, WANG S Q, LI D Y, et al. GeoLocator: a location-integrated Large Multimodal Model (LMM) for inferring geo-privacy [J]. *Applied Sciences*, 2024, 14(16): 7091.
- [35] JAYATI S, CHOI E, BURTON H, et al. Leveraging large multimodal models to augment image-based building damage assessment [C]//*Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. New York, USA: ACM Press, 2024: 79-85.
- [36] HAO X X, CHEN W, YAN Y B, et al. UrbanVLP: multi-granularity vision-language pretraining for urban socioeconomic indicator prediction [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2403.16831>.
- [37] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2010.11929>.
- [38] CHEN X L, FAN H Q, GIRSHICK R, et al. Improved baselines with momentum contrastive learning [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2003.04297>.
- [39] TOLSTIKHIN I O, HOUSBY N, KOLESNIKOV A, et al. MLP-Mixer: an all-MLP architecture for vision [J]. *Advances in Neural Information Processing Systems*, 2021, 34: 24261-24272.
- [40] AYUPOV S, CHIRKOVA N. Parameter-efficient finetuning of transformers for source code [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2212.05901>.
- [41] ZHANG Z, SABUNCU M. Generalized cross-entropy loss for training deep neural networks with noisy labels [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/1805.07836?context=cs>.
- [42] 杨冬菊, 黄俊涛. 基于大语言模型的中文科技文献标注方法 [J]. *计算机工程*, 2024, 50(9): 113-120.
- YANG D J, HUANG J T. Chinese scientific literature annotation method based on large language model [J]. *Computer Engineering*, 2024, 50(9): 113-120. (in Chinese)
- [43] YASEEN M. What is YOLOv9: an in-depth exploration of the internal features of the next-generation object detector [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2409.07813>.
- [44] KHANAM R, HUSSAIN M. YOLOv11: an overview of the key architectural enhancements [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2410.17725>.
- [45] WU Z, LIU X, GILITSCHENSKI I. EventCLIP: adapting CLIP for event-based object recognition [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2306.06354>.
- [46] CHEN Z, WANG W Y, CAO Y, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2412.05271>.
- [47] GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The llama 3 herd of models [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2407.21783>.
- [48] WANG P, BAI S, TAN S N, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2409.12191>.
- [49] AGRAWAL P, ANTONIAK S, BOU HANNA E, et al. Pixtral 12B [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2410.07073>.
- [50] TEAM G, GEORGIEV P, LEI V I, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2403.05530>.
- [51] CHIANG W L, ZHENG L M, SHENG Y, et al. Chatbot arena: an open platform for evaluating LLMs by human preference [EB/OL]. [2024-12-19]. <https://arxiv.org/abs/2403.04132>.

文字编辑 金胡考
栏目编辑 赖玉玲