中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)04-0987-17

论文引用格式:Ma Q P, Bi H S, Zhang Q and Zhao X F. 2026. Construction of the Chinese bar chart dataset and research on data extraction methods. Journal of Image and Graphics, 31(4):0987-1003(马秋平, 毕航烁, 张琪, 赵晓凡. 2026. 中文条形图表数据集构建及数据抽取方法. 中国图象图形学报, 31(4):0987-1003)[DOI:10.11834/jig.250299]

中文条形图表数据集构建及数据抽取方法

马秋平,毕航烁,张琪*,赵晓凡

中国人民公安大学信息网络安全学院,北京　100038

摘　要:**目的**　图表作为直观高效的信息呈现方式,在科研与商业分析中扮演着重要角色。然而,当无法直接访问其底层原始数据时,基于图表进行深入分析便面临显著挑战。图表数据抽取技术旨在克服这一障碍,通过从视觉化的图表中精确提取数据,为后续的复杂指标计算、图表类型转换等下游任务提供关键的数据基础。本研究构建了一个大规模中文条形图数据集,并分别实现基于规则与大模型微调的图表数据抽取方法,以提升中文图表数据逆向提取的准确性与鲁棒性。**方法**　本研究构建了包含58 712幅多种类型中文条形图及其对应数据表格的数据集,含垂直/水平/堆叠条形图、多角度文本旋转等复杂场景,并衍生出图表文本识别、图例检测等专项数据集,为中文图表理解任务提供了高质量、多样化的基准数据支持。同时,提出了两种基准模型:基于规则的图表数据抽取方法和基于大模型微调的数据抽取方法。最后,本文设计并实现了一个集成多模型的图表数据抽取与类型转换系统,以验证方法的实际应用潜力。**结果**　基于规则的方法在中文条形图上取得了最佳的性能(69.97%);而基于大模型微调的方法在DVQA(understanding data visualization via question answering)数据集上的性能显著超越了先进方法UniChart(a universal vision-language pretrained model for chart comprehension and reasoning)(24.53%)和DePlot(one-shot visual language reasoning by plot-to-table translation)(41.29%),分别高出36.75%和19.99%,表明了该方法在跨语言场景下的卓越泛化能力。实验表明,基于规则的方法展现出处理特定图表类型的最佳性能,尤其在处理复杂图表结构方面具有明显优势;而基于大模型微调的方法虽然在单一图表类型上表现略逊,但具备更强的泛化能力和鲁棒性。**结论**　本文创建的中文条形图表数据集为中文图表理解任务提供了高质量、多样化的基准数据支持,并设计了一个集成多模型的图表数据抽取与类型转换系统,以验证方法的实际应用潜力。数据集开源地址https://doi.org/10.57760/sciencedb.j00240.00052,相关代码开源地址https://github.com/maqiuping59/ChineseChartExtract。

关键词:大语言模型微调;多模态数据抽取;中文图表数据集;视觉—语言联合学习;数据可视化逆向工程

Construction of the Chinese bar chart dataset and research on data extraction methods

Ma Qiuping, Bi Hangshuo, Zhang Qi*, Zhao Xiaofan

School of Information and Cyber Security, People's Public Security University of China, Beijing　100038, China

Abstract: **Objective** Charts are a vital tool for information presentation in research and business analysis, offering a clear and intuitive understanding of complex data relationships and trends. However, the inability to access the underlying raw data of these visual representations creates a serious barrier for in-depth analysis and further data utilization. Chart data

收稿日期:2025-07-14;**修回日期**:2025-10-30;**预印本日期**:2025-11-08

*通信作者:张琪　qi.zhang@ppsuc.edu.cn

基金项目:中央高校基本科研业务费专项资金资助(2024JKF18)

Supported by: Fundamental Research Funds for the Central Universities (2024JKF18)

extraction (CDE) technology aims to bridge this gap by accurately extracting data from visual charts and converting them into structured numerical values or tables, enabling complex metric calculations, chart type conversions, and other downstream tasks. **Method** This study addresses the challenges of CDE in the Chinese language context. We present the construction of the large-scale Chinese bar chart dataset, which contains 58 712 images encompassing various types, including vertical, horizontal, and stacked bar charts, with complex scenarios (e. g. , multiangle text rotation). We derive specialized datasets for chart text recognition, text classification, and legend detection, providing a robust foundation for Chinese chart understanding tasks. To evaluate and compare different CDE approaches, we propose two benchmark models. Rule-based chart data extraction: this method combines text recognition differentiable binarization net (DBNet) and object detection you only look once version 5 (YOLOv5) algorithms to extract textual elements and graphical features. A rule-based library, which is designed based on chart structure characteristics, is then employed for data structure reconstruction. Large model fine-tuning for chart data extraction: this approach utilizes a pretrained large vision-language model (i. e. , Qwen-VL-2B) and adapts it to the CDE task through parameter-efficient fine-tuning low-rank adaptation. This method leverages the model's general knowledge representation capabilities and enhances its performance on specific tasks with minimal computational resources. To assess the performance of the proposed methods and existing state-of-the-art models, we design and implement a CDE and type conversion system by using PyQt5. This system integrates multiple models, including the rule-based approach, the fine-tuned large model, and other open-source CDE models, enabling users to easily extract data and convert chart types. **Result** Experiments conducted on the Chinese bar chart dataset demonstrate the effectiveness of the proposed methods. Among the compared methods, the rule-based approach achieves the best performance overall (F1 score of 69. 97%), particularly for charts without data labels. This success can be attributed to the utilization of optical character recognition and object detection models specifically trained for the dataset, along with the image segmentation and multiscale line sampling true value correction algorithm. On the English Understanding Data Visualization via Question Answering Dataset, the fine-tuned large model outperforms existing state-of-the-art methods, achieving an F1 score of 61. 28%. This result highlights the model's strong generalization capabilities and robustness in cross-lingual scenarios. Qualitative analysis further confirms the effectiveness of the proposed methods. In comparison with other models, our approach demonstrates superior performance in handling complex chart structures and irregular text, accurately extracting metadata from charts. Ablation experiments are also conducted to investigate the contributions of the different components of the large model fine-tuning approach. Results reveal that the combined fine-tuning of the visual encoder, language decoder, and cross-modal adapter achieves excellent performance, indicating the necessity of a holistic optimization strategy for complex CDE tasks. **Conclusion** This study presents a comprehensive approach to CDE in the Chinese language context. The construction of a large-scale Chinese bar chart dataset and the proposal of two benchmark models provide valuable resources and reference standards for future research. The developed CDE and type conversion system demonstrates the practical application potential of the proposed methods. The current dataset focuses on bar charts and synthetic data, so future work may explore the integration of real-world charts and additional chart types to enhance dataset realism and diversity. Further research should also be conducted to investigate other robust model structures and training methods to address the limitations of large models in managing complex chart structures and irregular text. The developed system can be further expanded with functionalities, such as supporting other chart type conversions and chart style modifications. In conclusion, this study provides a substantial contribution to the field of Chinese CDE by offering new solutions and promoting the development of multimodal data analysis. We believe that further research and development in this area will unlock the full potential of chart data and enable efficient and insightful data analyses in various domains. The dataset is linked at https://doi.org/10.57760/sciencedb.j00240.00052. The code is linked at https://github.com/maqiuping59/ChineseChartExtract.

Key words: large language model fine-tuning; multimodal data extraction; Chinese chart dataset; vision-language joint learning; data visualization reverse engineering

0 引　言

在信息爆炸的时代，图表作为一种直观、高效的信息载体，广泛应用于科研报告、商业分析、新闻报道等多个领域。图表能够将复杂的数据关系和趋势以更加直观的方式呈现出来，使得用户更容易理解数据之间的关系。然而，这也带来了新的挑战：图表往往将原始的、底层的数据“封装”在图形元素（如柱状图的高度、折线图的点坐标、饼图的扇区角度等）之中，用户难以直接访问和利用这些底层数值。当需要对这些数据进行深入分析或者进行更复杂的计算（如计算增长率、市场份额、相关性等）时，仅仅依赖对图表的视觉解读是远远不够的，甚至可能因主观判断误差而导致结论偏差。这种“数据可见但不可用”的困境，限制了数据价值的进一步挖掘和利用。

鉴于上述背景，如何从图表中自动、准确地抽取原始数据，成为连接视觉信息与结构化数据的关键桥梁。图表数据抽取（chart data extraction，CDE）技术的目的是通过计算机视觉和自然语言处理等手段，将数据从表格中准确地恢复出来，并将其转换为可供机器直接处理的结构化数值或表格形式。如图 1 所示，图表数据抽取不仅能够解决因无法访问底层数据而导致的分析困难，更能为一系列下游任务，如图表摘要生成（Obeid 和 Hoque，2020；Liu 等，2023b）、图表问答（黎颖 等，2023；Kahou 等，2018；Masry 等，2022；Wei 等，2024；Zhang 等，2024b）等提供数据来源。

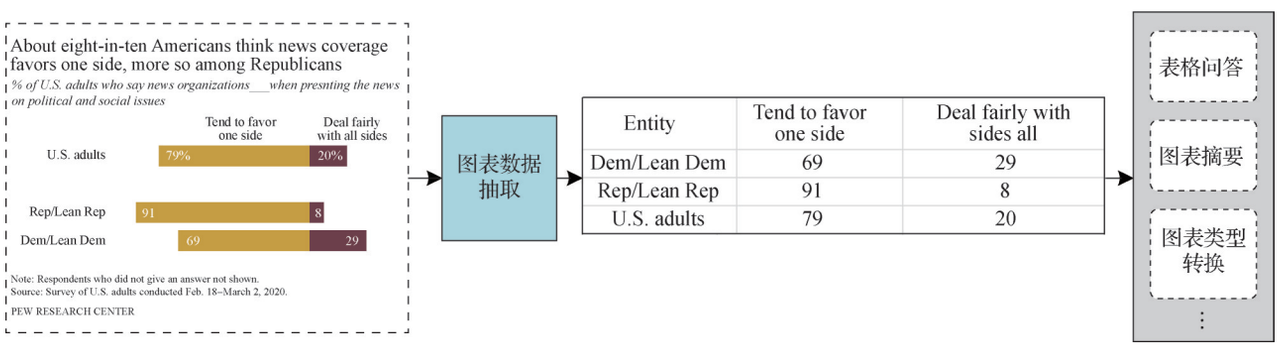


图 1　图表数据抽取

Fig. 1　Chart data extraction

早期的图表数据抽取方法，大多是依赖传统计算机视觉技术，如边缘检测、形状识别和光学字符识别（optical character recognition，OCR）等（Carderas 等，2020）来实现。例如，使用 Canny 边缘检测和 Hough 变换等检测图表的刻度轴等。然而，这些方法在处理复杂结构的图表时存在较大局限性（Ma 等，2021）。深度学习方法的引入显著提升了对图表复杂结构的解析能力（Ma 等，2021；Kanroo 等，2025）。此外，一些针对特定图表类型的抽取算法也为不同类型图表的自动化解析提供了多样化的技术路径（Carderas 等，2020；Lal 等，2023）。近年来，多模态大模型（multi-modal large model，MLLM）强大的视觉理解能力和推理能力使得图表数据抽取（Obeid 和 Hoque，2020；Zhang 等，2024a）迎来了新的突破。

目前图表数据抽取方法的研究围绕英文数据图表数据集展开，专门针对于中文图表的数据提取方法研究较少。中文图表在文本内容、布局方式和视觉风格上与英文图表存在显著差异，这为中文图表数据抽取带来了独特的挑战。为了解决这些问题，本文开展了以下研究：

1）构建了一个包含 58 712 幅多种类型中文条形图及其对应数据表格的数据集。在此基础上，进一步构建了图表文本 OCR 数据集、图表文本分类数据集和图例检测数据集，所有数据集均已开源，可从 https://doi.org/10.57760/sciencedb.j00240.00052 下载。

2）提出了两种基准模型：基于规则的图表数据抽取方法和基于大模型微调的数据抽取方法，相关代码开源至 https://github.com/maqiuping59/ChineseChartExtract。

3）为系统评估多模态大模型在图表数据抽取任务中的性能表现，创建了图表结构化评估模型，通过将表格提取为三元组形式，解决了现有评估方法在文本顺序敏感性和图表矩阵转置场景下的失效问题，显著提升了评估鲁棒性，其核心代码及评估框架

已开源至HuggingFace平台(访问地址:https://huggingface.co/spaces/maqiuping59/table_markdown)。

4)为进一步推动研究成果的实际应用,本文基于PyQt5开发了一个图表抽取及类型转换系统,该系统除本文提出的两种模型外,还集成了现有多种图表提取模型,能够轻松实现图表数据抽取和图表转换。

1 相关工作

1.1 图表数据集

图表多样性数据集的构建和针对该类型数据集的特定算法优化是图表数据集研究的一个方向,旨在增加科学图表数据集类型和数据多样性,能提供或开发出可处理各种多样性的算法。

FigureQA(an annotated figure dataset for visual reasoning)(Kahou等,2018)数据集是在探索科学图表视觉推理迈出的第一步,在FigureQA中包含超过10万张图表,同时在其中提出超过100万道由模板所生成的问题,问题包含了科学图表中可能存在的大多数推理(如最大/小值、曲线面积等)。与FigureQA相比,DVQA(understanding data visualization via question answering)(Kafle等,2018)更加关注于柱状图的解析,数据集采用程序化方式保证了数据集样式、文本和数据方面的多样性,其中DVQA包含300 000幅不同样式的条形图,并且给每个条形图都增加了元数据,这些元数据包括图表数据以及所有图形元素和文本元素的坐标。数值类型分为3类:1~10的线性数字、10%的百分数以及1~1 010的指数类型,所有的数字都是整数。

FigureQA和DVQA都使用了合成数据集,在数据上并不完全适应真实的图表,因此PlotQA(reasoning over scientific plots)(Methani等,2020)扩展了这一方向,其数据集来源于现实数据源,并利用众包方式基于问题模板生成问题,共包含28.9百万个问答对。

以上这些数据集的侧重点还是在柱状图的解析上,数据集最大的缺点是没有包含更多的图表类型。为了解决上述问题,Bajić和Job(2022)提出了更丰富的针对特殊图表的数据集和方法:一个包括饼图、环形图、日晷图以及华夫图表在内的多达120 K幅图像数据集,同时他们的方法也是基于二值图像处理的方式来提高通用性。

1.2 图表数据抽取

图表中存在两类元素,一类是文本元素:标题、图例、刻度等,主要用于对图表的解释说明;另一类是对数值关系进行表示的图形元素,如饼图中的圆心角度数表示数值大小,折线图中的折线斜率变化来表达数值的变化。早期的图表数据抽取依赖于传统的计算机视觉技术,例如Carderas等人(2020)提出的自动化数据提取流程使用传统的图像处理方式(Otsu二值化,Canny边缘检测)完成柱状图的数据提取,但其主要误差来源于OCR对数字的误读问题。这暴露了基于传统方法的显著局限性,即对复杂图像特征的处理能力不足。

随着深度学习的兴起,基于深度学习的框架逐渐成为图表数据抽取领域的主流研究方向。这些方法通过端到端的架构有效地克服了传统方法中的诸多瓶颈。例如,Ma等人(2021)提出的框架以ResNet-50(He等,2016)和R-CNN(region-convolutional neural network)(Girshick等,2014)为主干,通过多层级回归头和特征融合显著提升了框检测性能,特别是在高交并比(intersection over union,IoU)值下的表现。此外,其点检测模块通过分割技术生成热图掩膜,能够有效处理背景噪声,并在密集点检测中表现优异。然而,深度学习方法的广泛应用也带来了新的挑战:训练复杂深度学习框架需要大量标注数据,而现有公开数据集的种类和规模尚难以满足这一需求。因此,Bajić和Job(2022)提出的大规模合成数据集为未来研究提供了宝贵的资源。

基于关键点检测的方法在图表数据抽取领域崭露头角。这类方法利用关键点检测网络(如HRNet(high-resolution net)(Sun等,2019))识别图表中的关键点,例如柱状图的柱顶点、折线图的折点等。通过更精确地定位图表中的数据点,为数据提取提供了新的思路。基于目标检测的方法则将图表元素视为目标,利用目标检测网络进行识别和定位。这类方法通常具有较高的检测精度和速度,能够有效地处理多种类型的图表。

1.3 大模型微调

近年来,大语言模型(large language model,LLM)在自然语言处理任务,如机器翻译、摘要生成以及智能问答等上取得了惊人的表现,然而,随着大模型技术的不断发展,大模型的参数规模也迎来了爆发式增长。在对大模型适配下游任务时,如果对

大模型全部参数进行训练，则需要巨大的数据量和运算成本。在此背景下，参数高效微调技术PEFT（parameter-efficient fine tuning）（Houlsby等，2019）应运而生。参数高效微调的原理是仅对一小部分自身参数或者外部引入参数进行微调，从而为模型带来整个性能的变化。由于只训练一小部分参数，PEFT技术极大程度地降低了训练大模型的算力需求。

Lialin等人（2024）按照方法类型将PEFT分为3类，分别为加法（addition-based）、选择（selection-based）和重新参数化（reparametrization-based）。

基于加法的方法是一种通过引入附加参数矩阵或模块来适应特定下游任务的轻量级调整策略。该方法的核心思想在于保留预训练模型的原始权重不变，只训练新引入部分的参数，实现对下游任务的适配。常见的基于加法的方法有适配器（adapter）（He等，2022；Pfeiffer等，2020）、软提示（soft prompts）（Lester等，2021；Liu等，2022）等。

基于选择的方法是一种通过动态选择并优化子集参数来适应特定下游任务的轻量级调整策略。该方法的核心思想在于并非对预训练模型的全部参数进行更新，而是基于某种选择机制（如基于梯度的选择、基于任务相关性的选择等）（Sung等，2021）识别出与下游任务最相关的参数子集，并在微调过程中仅针对该子集进行优化。这种方法有效降低了需要更新的参数规模，从而减少了计算资源消耗和过拟合风险，同时由于保留了大部分预训练参数，能够在一定程度上缓解灾难性遗忘问题，代表性方法为BitFit（bias-term fine-tuning）（Ben Zaken等，2022）。

基于重新参数化的方法通过引入可学习的低维参数空间对预训练模型的原始参数进行显式转换，从而实现高效的任务适应。这类方法的核心在于构建一个参数转换函数，该函数通常依赖于预训练模型参数和一个独立的、可学习的紧凑参数集，例如通过线性变换、门控机制或函数分解等方式对原始参数进行重新表达。最著名的基于重新参数化的方法是LoRA（low-rank adaptation）（Hu等，2022），该方法的核心思想在于将模型关键层的权重矩阵 $\boldsymbol{W}$ 分解为两个低秩矩阵 $\boldsymbol{W}_A$ 和 $\boldsymbol{W}_B$ 的乘积，即

$$\boldsymbol{W}' = \boldsymbol{W} + \boldsymbol{W}_A \times \boldsymbol{W}_B \tag{1}$$

式中，$\boldsymbol{W}_A$ 和 $\boldsymbol{W}_B$ 的秩远小于原始权重矩阵 $\boldsymbol{W}$ 的维度。在微调过程中，仅对低秩矩阵 $\boldsymbol{W}_A$ 和 $\boldsymbol{W}_B$ 进行优化，而保持原始权重矩阵 $\boldsymbol{W}$ 不变。

2 本文方法

为了全面且深入地评估数据抽取方法在中文图表数据集上的性能和潜力，首先构建一个具有代表性的中文图表数据集。基于该中文图表数据集，本文提出了两种具有代表性的基准方法，第1种是基于规则的图表数据抽取方法，该方法通过结合文本识别算法DBNet（differentiable binarization net）（Liao等，2020）与YOLOv5（you only look once version 5）目标检测算法实现文本元素提取，结合基于图表结构特征设计的规则库完成数据结构化重建；第2种为基于大模型微调的抽取方法，采用预训练模型作为基础架构，通过领域自适应微调策略，充分迁移模型在通用领域的知识表征能力，从而提升数据抽取的准确性与鲁棒性。

2.1 中文图表数据集构建

当前，在图表数据集领域，已有的资源大多以英文内容为主，开源的中文图表数据集几乎没有，这为中文图表相关的研究和应用带来了一定的局限性。本文借鉴DVQA（Kafle等，2018）数据集的构建方法，创建了一个专注于中文环境的图表数据集。为了确保数据集的真实性和实用性，首先参考权威的国家统计局网站，从中选取了在实际应用中广泛使用的24个数据标签类别，共计262个具体的标签。这些标签类别涵盖了社会经济、人口统计以及产业发展等多个重要领域，部分典型的标签类别在表1中列出。

为了进一步增强数据集的多样性和实用性，如表2所示，本文设置了10种不同的数值维度。这些数值维度不仅提供了丰富的数值范围，还包含了多种取值类型，从而能够模拟真实应用场景中可能遇到的各种数据分布和变化情况。

该数据集精心设计了多种类型的中文条形图表，旨在涵盖实际应用中可能遇到的各种情况。具体而言，数据集不仅包括了常规的垂直条形图和水平条形图，还特别引入了更具挑战性的堆叠条形图，以测试方法在不同复杂度图表上的表现。此外，为了进一步增加数据集的多样性和实用性，为每种图表类型设置了多样化的属性标签。这些属性标签包括但不限于是否带有数据标签、文本是否旋转45°、90°等，这些细节的加入使得数据集更加贴近真实应用场景，同时也对数据抽取方法提出了更高的要求。

表1 标签示例

Table 1 Examples of labels

标签类别	标签
性别	男、女
年龄	0–10岁、10–20岁、20–30岁、30–40岁、40–50岁、50–60岁、60–70岁、70–80岁、80–90岁、90–100岁
季节	春、夏、秋、冬
省份	北京、天津、河北、山西、内蒙古、辽宁、吉林、黑龙江、上海、江苏、浙江、安徽、福建、江西、山东、河南、湖北、湖南、广东、广西、海南、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏、新疆、香港、澳门、台湾
月份	1月、2月、3月、4月、5月、6月、7月、8月、9月、10月、11月、12月
政治面貌	党员、群众、团员、其他
婚姻状态	未婚、已婚、离异、丧偶

表2 数值维度类别示例

Table 2 Numeric dimension category examples

数值维度	数值范围	数值类别
年龄/岁	0～100	整数
收入/元	0～5 000 000	整数
身高/cm	0～240	小数
体温/℃	35～42	小数
人数/人	0～1 000	整数
比例	0～1	小数
增长率/%	−10～10	小数
天数/天	0～365	整数
体重/kg	0～200	小数
心率/次	0～200	整数

除了图表本身，数据集还为每张图表提供了与之对应的数据表格和标题文本，这些信息对于理解图表内容、验证抽取结果的准确性至关重要。

本文选择Python编程语言中最受欢迎且应用广泛的数据可视化库——Matplotlib，来负责生成研究所需的图表图像。Matplotlib以其丰富的功能、灵活的配置选项以及出色的兼容性，成为数据科学家和研究人员在数据可视化任务中的首选工具。通过利用Matplotlib库，能够精确地控制图表的每一个细节，从数据点的绘制到坐标轴的标注，从图例的添加到标题的设定，从而确保生成的图表图像不仅符合研究的需求，而且在视觉上具有高度的可读性和吸引力。

数据集的构建过程如算法1所示。首先，从预先定义的标签类别库中随机选取两个类别，分别作为X轴标签（以下简称“标签”）和图例标签（以下简称“图例”）。随后，从选定的标签类别的具体标签中随机选择2～10个作为X轴标签值，从图例类别的具体标签中随机选择1～5个作为图例项。至此，图表所包含的数据点的规模得以确定。接下来，从预先定义的数值维度类别中随机选择一个类别，并根据该类别预先设定的数值范围生成一系列随机数，从而构建数据表。同时，根据固定的模板生成图表标题，其格式为“不同 <标签类别> 下的 <图例类别> 的 <数值维度>”。最后，利用Matplotlib库生成具有不同视觉风格的图表。

数据生成完成后，为确保数据集的质量，本文对其进行了进一步筛选。删除了存在条形重合且文本标签大部分被遮挡、影响阅读的样本。数据集包含垂直、水平、堆叠、带数值标签、标签文本旋转45°和90°等类型条形图共58 712幅。数据集示例如图2所示。

算法1：图表数据集合成算法。

输入：标签类型集合Labels、数值维度类别Tricks。

输出：数据表格Table、图表Chart、标题Title。

```
while idx < num:
    x_label ← RandomChoice(Labels,1);
    legend ← RandomChoice(Labels,1);
    range ← RandomChoice(Tricks,1);
    label_num ← Randint(2,10);
    legend_num ← Randint(1,5);
    Table = [];
    for (row = 0; row < label_num; row++):
        for (col = 0; col < legend_num; col++):
            Table[row][col] = Random(range. low, range.high);
    Title ← “不同{x_label}下的{legend}的{range}”;
    Chart ← Plot(Table);
return: Table, Chart, Title。
```

对于每一类别，按照 7∶2∶1 的比例将数据集划分为训练集、验证集和测试集。数据集的最终统计数据如表 3 所示。与 DVQA 数据集相比，本文提出的数据集图表文本元素更加丰富、数值范围更贴近实际。

2.2 基于规则的图表数据抽取方法

与普通场景下的 OCR 任务不同，针对图表的 OCR 任务需要处理多种布局的文本、复杂的背景、多方向的字符以及文本与图表元素的关联性，此外，图表中的线条、颜色块和图形元素会干扰文本定位和识别，因此，针对普通场景下的 OCR 方法在图表字符识别中表现不佳。

本文提出的基于规则的图表数据抽取算法如图 3 所示。使用 OCR 算法对图表中的文本信息进行检测，然后对返回的文本位置信息使用图表的长和宽进行归一化操作，接着使用随机森林算法（random forest algorithm）对图表检测到的文本进行分类，文本类别包含标题、横轴标题、纵轴标题、图例、标签和数值刻度 6 类。同时，使用 YOLOv5n 对图表中的图

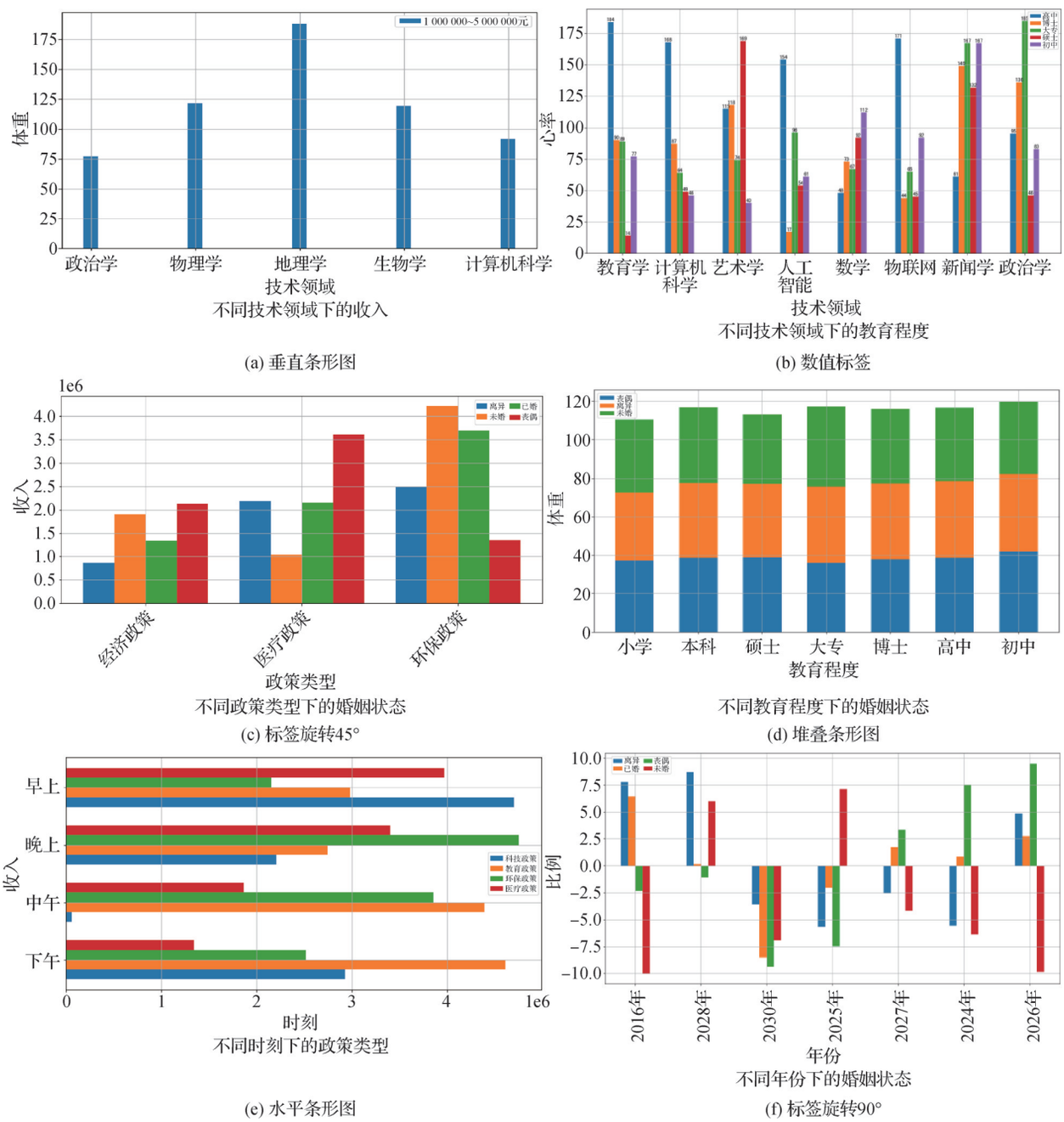


图 2　条形图示例

Fig. 2　Example bar chart images ((a) vertical bar chart; (b) with value label; (c) trick with 45° rotation; (d) stacked bar chart; (e) horizontal bar chart; (f) trick with 90° rotation)

表3　数据集中不同条形图类型统计信息

Table 3　Statistics for different types of bars in the dataset

/幅

划分	垂直条形图	水平条形图	堆叠条形图	数据标签	标签旋转45°	标签旋转90°	合计
训练集	6 884	6 946	6 920	6 549	6 982	6 814	41 095
验证集	1 967	1 984	1 977	1 871	1 995	1 947	11 741
测试集	984	994	990	936	998	974	5 876
合计	9 835	9 924	9 887	9 356	9 975	9 735	58 712

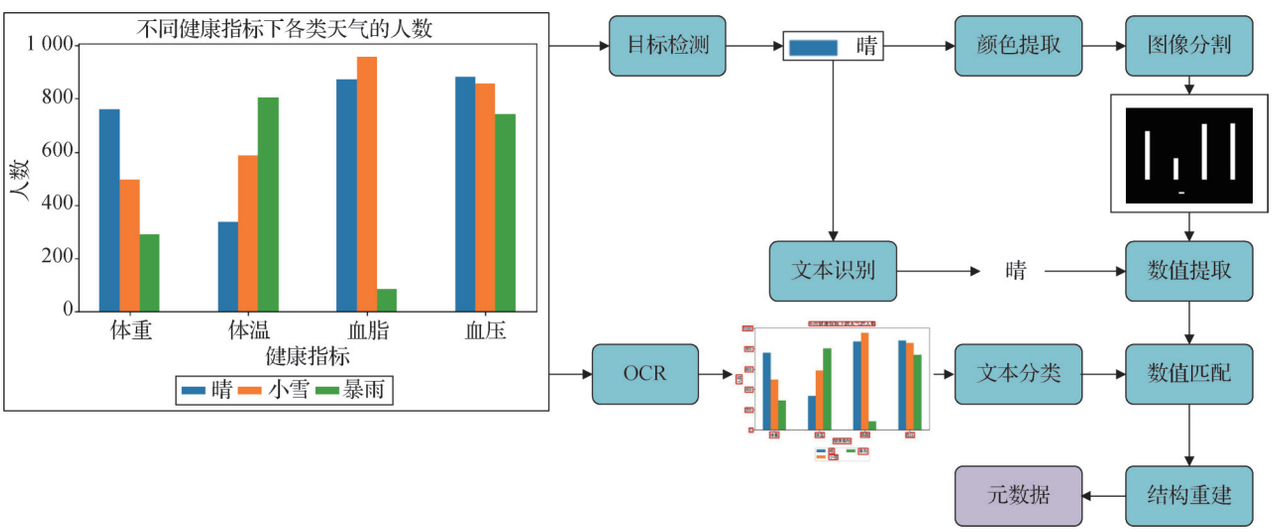


图3　基于规则的图表数据抽取方法

Fig. 3　Rule-based data extraction method for charts

例进行检测，针对检测到的图例，分别使用字符识别和OpenCV获得图例文本和颜色特征。接着使用提取到的颜色特征对每一类图例进行分割，得到对应条形图的位置信息。

在对分割后的条形图提取其数值的过程中，受文本检测误差影响，基于位置推导的数值与真实值存在系统性偏差，导致数值还原结果与真实值产生显著偏差。为了尽可能减少这种误差的影响，提出基于多刻度线采样的真值校正算法，如图4所示，首先在OCR检测结果中随机获取4条刻度线(T_{ci}, H_{ci})（图4蓝色虚线标注），T_{ci}代表第i条刻度的刻度值，H_{ci}代表文本边界框中心点的纵坐标（如果为水平条形图，则代表边界框的中心点的横坐标），假设需要确定数值的条形的纵坐标为H_t，则其数值V_t的计算为

$$V_t = \frac{1}{3}\sum_{0\leqslant i<j\leqslant 3}\left(\frac{T_{cj}-T_{ci}}{H_{cj}-H_{ci}}\times\left(H_t-H_{cj}\right)+T_{cj}\right) \tag{2}$$

数值匹配的难点在于精确识别标签与数值之间

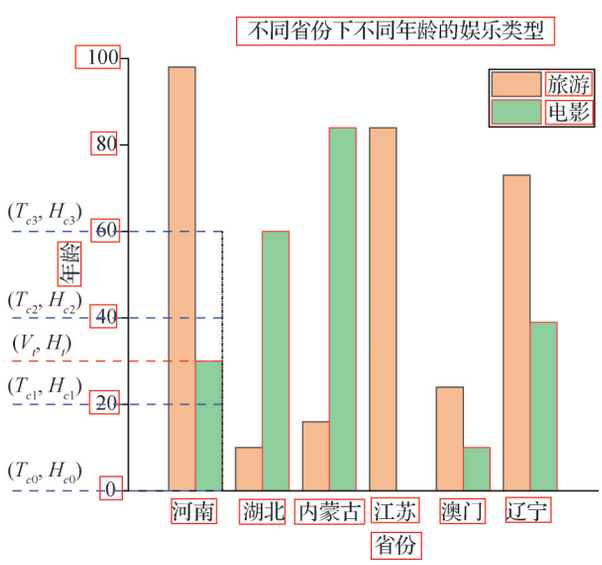


图4　多刻度线采样的真值校正

Fig. 4　True value correction of multi-scale line sampling

的对应关系，由于条形图中可能存在值为零的情况，因此不能简单地通过条形和标签的位置顺序进行匹配。为了实现标签与数值的准确匹配，使用基于距

离的匹配方法。如图 5 所示，首先计算条形中心点与每个标签中心点的欧氏距离

$$d_i = \sqrt{\left(x_b - x_i\right)^2 + \left(y_b - y_i\right)^2} \tag{3}$$

式中，x_b、y_b分别代表条形图中心点的横纵坐标，x_i、y_i分别代表第i个标签的中心点的横纵坐标，d_i代表条形图与第i个标签中心点之间的距离，距离最小的标签即为当前条形图所属标签。

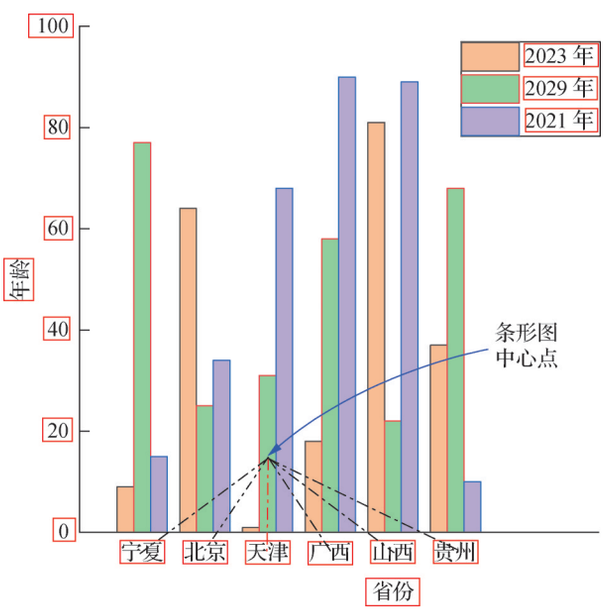


图 5　数值—标签匹配过程

Fig. 5　Numerical-label matching process

2.3　基于 Qwen-VL 进行 LoRA 微调的图表数据抽取方法

2.3.1　Qwen-VL 系列模型综述

大型视觉语言模型（large vision-language model，LVLM）在传统大语言模型的基础上引入视觉感知能力，实现了对图像等多模态信息的理解和处理。Qwen-VL（Bai 等，2023b）以大语言模型 Qwen-7B（Bai 等，2023a）作为基础组件，在此基础上接入视觉分支。Qwen-VL 的视觉编码器使用 ViT（vision Transformer）（Dosovitskiy 等，2021）架构，输入 ViT 的图像首先被分割成步幅为 14 的块（patch），从而生成一组图像特征。为了缓解长图像特征序列带来的效率问题，Qwen-VL 引入了一种视觉语言适配器，可以压缩图像特征。该适配器包括一个随机初始化的单层交叉注意力（cross-attention）模块。该模块使用一组可训练向量（embeddings）作为查询（query）向量，并使用来自可视编码器的图像特征作为 Cross-Attention 的键值（key）。此机制将视觉特征序列压缩为固定长度 256，随后将长度为 256 的压缩图像特征序列输入到大型语言模型中。

Qwen2-VL（Wang 等，2024）在 Qwen-VL 的基础上引入了 NaViT（native resolution ViT）（Dehghani 等，2023），这使得 Qwen2-VL 可以处理任何分辨率的图像，并将它们动态转换为可变数量的视觉标记（tokens）。Qwen2. 5-VL（Bai 等，2025）的视觉编码器采用了重新设计的 ViT 视觉转换器架构，并结合了 2D-RoPE 和窗口注意力来支持原生输入分辨率，同时加速整个视觉编码器的计算。在训练和推理过程中，输入图像的高度和宽度在输入到 ViT 之前被调整为 28 的倍数。视觉编码器通过将图像分割成步幅为 14 的补丁来处理图像，从而生成一组图像特征。为了解决长序列图像特征带来的效率挑战，Qwen2. 5-VL 采用了一种简单而有效的方法，在将特征序列输入到大型语言模型之前使用一个多层感知机对其进行压缩，压缩后的维度与输入到 LLM 中的文本的维度一致。这种方法不仅降低了计算成本，而且还提供了一种灵活的方法来动态压缩不同长度的图像特征序列。

2.3.2　指令微调数据集构建

现有的图表问答数据集只提供了图表和图表元数据，不能直接用于大模型微调，因此基于本文提出的中文图表数据集和图表问答数据集 DVQA 构建了大模型指令数据集，数据集形式如图 6 所示。为了便于大模型读取和输出表格，采用 markdown 形式表示数据表格。markdown 是一种轻量级标记语言，它允许人们使用易于阅读、易于编写的纯文本格式编写文档。与 HTML 使用各种标签来表示表格相比，markdown 使用“|”来划分单元格，使用“|---|”来区分标题行。

2.3.3　LoRA 微调

受限于实验室环境配置，本文选用 Qwen-VL-2B 模型进行微调。Qwen-VL 模型如图 7 所示，本文对模型中的视觉编码器、视觉适配器和大模型部分进行 LoRA 微调，微调过程中，将输入图片尺寸统一预处理为 224 × 224 像素，分别使用高斯分布和零初始化权重两个低秩矩阵的权重。

在自然语言处理领域，常用的生成文本评估指标包括 BLEU（bilingual evaluation understudy）（Papineni 等，2002）、ROUGE（recall-oriented understudy

```
[
  {
    "id": "bar_001",
    "conversations": [
      {
        "role": "human",
        "value": <image path>
      },
      {
        "role": "assistant",
        "value": " | tape | fury | sand | cold | bed |
              |--|--|--|--|--|
              | coach | 1000000000 | 1000000000 | 100000000 | 100000 | 1000000000 |
              | vision | 10 | 100000 | 1000000000 | 1000000000 | 1000000000 |
              | town | 1000000000 | 10000000 | 100000 | 1000 | 10000 |"
      }
    ]
  },
]
```


图6　指令微调数据集示例

Fig. 6　Examples of the instruction fine-tuning dataset

for gisting evaluation)(Lin, 2004)和METEOR(metric for evaluation of translation with explicit ordering)(Banerjee和Lavie, 2005)。这些指标通过比较大模型预测文本与真实文本之间的相似度,能够有效评估生成文本的质量。然而,这些指标在评估markdown格式表格数据的准确率方面存在局限性。具体而言,标签和数值的同步变化虽然不会影响表格的实际准确率,但却会导致上述指标的计算结果发生变化。为了更好地评估基于多模态大模型的抽取效果,本文创建了针对markdown格式数据表格的评估模型,首先通过正则表达式从模型输出中提取表格,接着将每个单元格表示为如下的三元组形式

$$Cell_{ij}=\left\{\boldsymbol{set}\left(R_i,C_j\right):V_{ij}\right\} \tag{4}$$

式中,$\boldsymbol{set}$代表无序集合,V_{ij}代表第i行第j列的数值,R_i代表第i行行名,C_j代表第j列列名,由于采用无序集合表示数值所在行和列,因此可以避免表格转置对评估指标的影响。本文参考Carderas等人(2020)、Liu等人(2023a)、Han等人(2023)研究结果中的标准,对于文本元素要求精确匹配,对于数值允许与真实值之间存在5%的误差,对于预测值$\widehat{Cell}_{mn}$和真实值$Cell_{ij}$,判断预测是否准确的具体计算为

$$\begin{cases}0 & \boldsymbol{set}(R_i,C_j)\neq\widehat{\boldsymbol{set}(R_m,C_n)}\ \text{or}\ \left|\dfrac{V_{ij}-\hat{V}_{mn}}{V_{ij}}\right|>5\% \\ 1 & \boldsymbol{set}(R_i,C_j)=\widehat{\boldsymbol{set}(R_m,C_n)}\ \text{and}\ \left|\dfrac{V_{ij}-\hat{V}_{mn}}{V_{ij}}\right|\leqslant 5\%\end{cases} \tag{5}$$

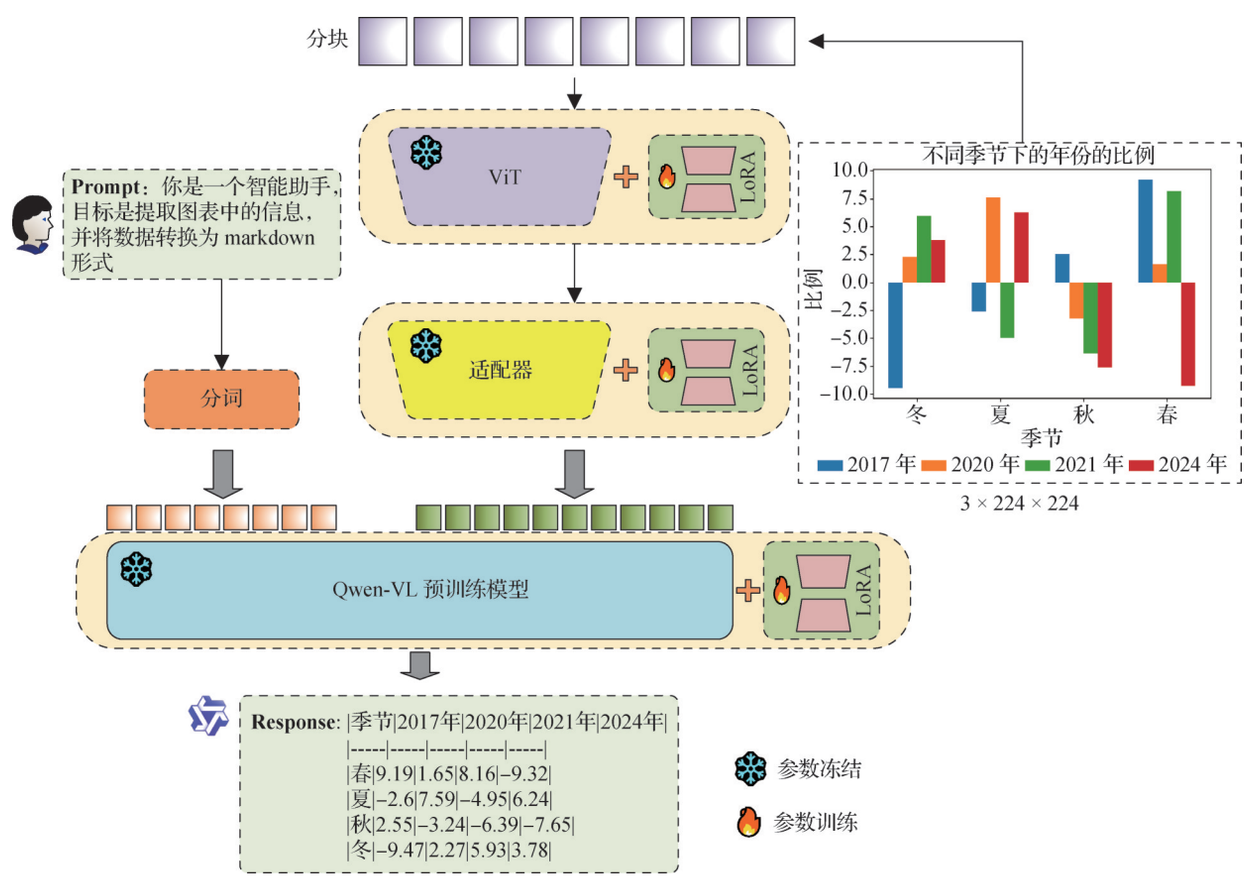


图7　Qwen-VL微调框架图

Fig. 7　Qwen-VL fine-tuning framework diagram

3 实验及结果分析

3.1 数据集

为了全面评估所提方法在中文和英文数据集上的性能表现，除本文构建的中文图表数据集外，还从英文图表数据集中随机抽取了2 500幅各类条形图。针对这些英文图表，利用标注工具完成了图表文本检测、文本识别、图例检测和文本分类等任务的数据标注工作，以用于训练和评估基于规则的数据抽取方法。随后，将所有数据集统一按照7∶2∶1的比例划分为训练集、验证集和测试集，具体划分情况详见表5。

表5 数据集划分情况

Table 5 Division of the dataset

/幅

	训练集	验证集	测试集	合计
图表文本检测	1 292	369	184	1 845
图表文本识别	20 800	5 943	2 971	29 714
图表图例检测	581	246	255	1 082
图表文本分类	1 292	369	184	1 845
Qwen2.5-VL微调（DVQA数据集）	1 500	500	500	2 500

3.2 评估指标

3.2.1 文本识别评估指标

在文本识别任务中，评估指标的选择直接影响对模型性能的客观评价。现有的评估方法中，编辑距离（edit distanc）和字符准确率（character accuracy）是常用的指标，但在特定场景下存在明显的局限性。编辑距离通过计算两个字符串之间的最小操作次数来衡量其相似性，然而对于短文本（如标签或数值），其区分能力有限。例如，标签“2019和“2020”的编辑距离仅为1，但语义差异显著；字符准确率则无法有效评估数值误差，如“100”与“1000”的字符差异虽小，数值误差却较大。

针对这些局限性，本文选择准确率（accuracy）作为核心评估指标。准确率的定义为预测文本与真实文本完全匹配的比例，其优势在于能够直接反映模型在严格匹配条件下的性能表现。

3.2.2 图表数据抽取评估指标

在图表数据抽取任务中，为了评估模型的性能，本文使用精确率（precision）、召回率（recall）和F1-score作为主要的评估指标。

3.3 实验结果

图表文本检测与识别结果如表6所示，在中文图表数据集上的文本检测精确率为98.5%，文本识别的准确率为98.74%，在英文图表数据集DVQA上的文本检测的精确率为99.24%，文本识别的准确率为99.53%。

表6 图表文本检测和识别结果

Table 6 Results of the chart text detection and recognition

数据集	文本检测			文本识别
	precision	recall	F1-score	accuracy
中文图表	0.985 0	0.911 7	0.946 9	0.987 4
DVQA	0.992 4	0.943 5	0.952 6	0.995 3

对于图表文本分类，本文以图表方向以及归一化后的文本位置为特征，使用随机森林算法进行分类，分类结果如表7所示，可以发现，对所有类型的图表文本分类精确率均达到了96%以上。

表7 图表文本分类结果

Table 7 Results of chart text classification

	precision	recall	F1-score	样条数
标题	1.00	1.00	1.00	249
横轴标题	0.96	0.95	0.95	248
纵轴标题	0.99	0.98	0.99	285
标签	0.99	0.98	0.99	1 176
图例	0.99	1.00	1.00	730
刻度	0.99	1.00	1.00	2 139
总体	0.99	0.99	0.99	4 827

使用Yolov5n识别图例实验中设置最大训练轮次设置为200，最优结果出现在188轮，此时在测试集上的准确率为99.39%。

中文图表数据集上的实验结果如表8所示。本文基于规则的图表数据抽取方法在除带有数据标签的条形图以外的其他图表类型上都取得了最好的效果。原因是本文利用针对数据集训练过的OCR模型和目标检测模型来抽取文本元素和图形元素。此

外，基于图像分割和基于多刻度线采样的真值校正算法也使得基于规则的图表抽取算法的数值抽取性能得到了提升。而对于带有数据标签的条形图来说，由于有数据标签的存在干扰了文本分类的结果，因此该方法在该种图表上的效果较差。

对比没有微调的模型，经微调后的模型无论是哪种类型的条形图都有明显的改善，准确率、召回率以及F1-score分别提升了56.66%，57.78%和58.95%。从结果来看，与基于规则的方法刚好相反，多模态大模型的方法在带有标签的条形图上获得了较好的结果，说明大模型有着优秀的文本抽取能力；但相对于垂直条形图来说，经微调后的模型在文本标签旋转45°时F1-score下降了30.88%，旋转90°的时候下降了45.42%，说明大模型在不规则文本识别上的性能还有待提高。另外，大模型在堆叠条形图上的效果也非常不好，这也说明大模型对复杂的图形元素排列有待提高。

为进一步验证微调后的模型对英文数据集的抽取能力，本文从DVQA数据集中随机抽取了2 500幅图表进行实验。同时，选取了当前先进的可以将输出图表形式准确转换为markdown形式的图表抽取模型Unichart（Masry等，2023）和DePlot（Liu等，2023a）进行对比实验，实验结果如表9所示。微调后的模型在DVQA数据集上的F1-score达到了61.28%，超越了UniChart和DePlot，取得了最优性能。

表8　中文图表数据集实验结果

Table 8　Experimental results on Chinese chart dataset

/%

		Qwen-VL	Rule-based	Qwen-VL-LoRA
垂直条形图	precision	3.38	**92.36**	78.73
	recall	6.00	**93.14**	79.63
	F1-score	4.32	**92.75**	79.18
水平条形图	precision	9.26	**89.25**	72.91
	recall	11.23	**91.33**	73.01
	F1-score	10.15	**90.28**	72.96
堆叠条形图	precision	0.81	**74.85**	38.44
	recall	1.07	**75.69**	38.71
	F1-score	0.92	**75.27**	38.57
标签旋转45°	precision	0.69	**56.89**	47.85
	recall	2.74	**58.24**	48.77
	F1-score	1.10	**57.56**	48.30
标签旋转90°	precision	0.00	**48.46**	32.94
	recall	0.00	**47.29**	34.62
	F1-score	0.00	**47.87**	33.76
数据标签	precision	15.24	12.25	**77.76**
	recall	17.27	11.39	**78.32**
	F1-score	16.19	11.32	**78.04**
整体	precision	0.83	**69.57**	57.49
	recall	5.87	**70.38**	63.65
	F1-score	1.46	**69.97**	60.41

注：加粗字体为每行最优结果。

表9　英文数据集实验结果

Table 9　Experimental results on English dataset

/%

方法	DVQA		
	precision	recall	F1-score
UniChart(Masry等，2023)	27.59	22.08	24.53
DePlot(Liu等，2023a)	46.53	37.11	41.29
Rule-base	30.25	31.36	30.80
Qwen-VL-LoRA	**60.70**	**61.87**	**61.28**

注：加粗字体为每列最优结果。

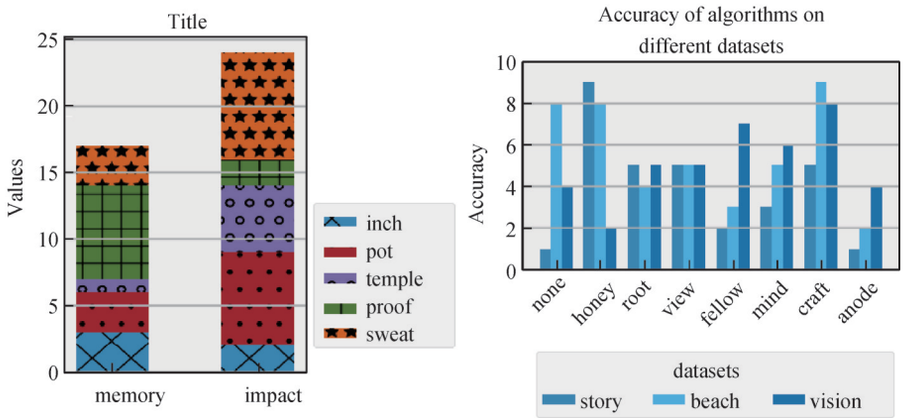
3.4　定性分析

本文在英文图表数据集DVQA上开展定性分析实验，可以直观地对不同模型的输出进行系统性比较。实验结果如图8和表10所示，在处理堆叠条形图图8（a）时，UniChart（Masry等，2023）出现了乱码现象，Chart-to-Table（Huang等，2024）出现了大模型幻觉现象，输出了与当前图表无关的内容，而DePlot（Liu等，2023a）模型虽然能够正确构建图表结构，但无法准确判断数值。在处理标签旋转图表图8（b）时，UniChart和Chart-to-Table均出现了无法准确识别文本的现象，例如，UniChart将“honey”错误地识别为“money”，Chart-to-Table将“craft”识别为“carf”。相比之下，DePlot模型虽然可以准确地识别出文本标签，但依旧无法准确识别出数值信息。在上述两种情况中，本文方法均能准确地从图表中恢复出元数据。

3.5　消融实验

3.5.1　文本分类

在基于随机森林的图表文本分类实验中，除文



(a) 堆叠条形图表　　(b) 标签旋转图表

图 8　定性分析实验图表示例

Fig. 8　Example of a qualitative analysis experimental chart((a)stacked bar chart;(b)label rotation chart)

表 10　定性分析结果

Table 10　Qualitative analysis results

方法	提示词		
UniChart (Masry 等,2023)	"<*extract_data_table*><*s_answer*>"	Characteristic | [unnamed data series #0] | [unnamed data series #1] | [unnamed data series #2] | [unnamed data series #2] | [unnamed data series #2] | [unnamed data series #3] | [unnamed data series #2] | [unnamed data series #2] | [unnamed data series #2]	Characteristic | story | beach | vision & none | 1.0 | 4.0 | 2.0 & money | 6.0 | 4.0 | 2.0 & foot | 5.0 | 4.0 | 2.0 & view | 5.0 | 4.0 | 2.0 & fellow | 3.0 | 4.0 | 2.0 & Mind | 5.0 | 4.0 | 2.0 & craft | 5.0 | 4.0 | 2.0 & anode | 2.0 | 4.0 | 2.0
Chart-to-Table (Huang 等,2024)	"<*data_table_generation*><*s_answer*>"	TITLE | TikTok usage frequency among internet users in the United States as of March 2020 &&& Entity | Values &&& 0.0, 0 – 0 | – &&& 0.0, 0 | – &&& 0.0, 0 | – &&& 0.0, 0 | – &&& 0.0, 0 | – &&& 0.0, 0 | – &&& 0.0, 0 | – &&& 0.0, 0 | – &&& 0.0, 0 | 24 &&& 0.0, 0 | 24 &&& 0.0, 0 | 24 &&& 0.0, 0 | 24 &&& 0.0, 0 | 24 &&& 0.0, 0 | 17 &&& 0.0, 0 | 31 &&& 0.0, 0 | 24 &&& 0.0, 0 | 17 &&& 0.0, 0 | 31 &&& 0.0, 0 | 24 &&& | None	TITLE | Accuracy of algorithms on different datasets &&& datasets | accuracy &&& 0.02 | video &&& 0.98 | video &&& 2.01 | fifteen &&& 3 | twenty &&& 2.62 | fifteen &&& 8.0 | black &&& 9 | carf &&& 6.0 | fifteen &&& 5.0 | black &&& 3.08 | fifteen &&& 9.0 | carf &&& 1.0 | sqloye &&& 9. | retail &&& 1.0 | retail &&& | None
DePlot (Liu 等，2023a)	*Generate underlying data table of the figure below:*	TITLE | Title <0x0A> Values | inch | pot | temple | proof | sweat <0x0A> memory | 2.5 | 9 | 9 | 14 | 17 <0x0A> impact | 2.5 | 9 | 9 | 8 | 14	TITLE | Accuracy of algorithms on different datasets <0x0A>Accuracy | story | beach | vision<0x0A>none | 4.01 | 8.22 | 3.98<0x0A>honey | 9.07 | 8.98 | 4.01 | 5.04<0x0A>root | 5.01 | 0.00 | 0.00 | 5.03<0x0A>view | 5.01 | 0.00 | 0.00 | 5.01<0x0A>fellow | 2.01 | 0.00 | 0.00 | 7.08<0x0A>mind | 3.03 | 0.00 | 0.00 | 6.02<0x0A>craft | 5.01 | 9.08 | 4.01 | 8.08<0x0A>anode | 1.00 | 0.00 | 0.00 | 4.00
本文	你是一个智能助手，目的是从图表中抽取真实数据，并转为 *markdown* 形式。	||memory|impact||--|--||inch|3|2||pot|3|7||temple|1|5||proof|7|2||sweat|3|8|	||none|honey|root|view|fellow|mind|craft|anode||--|--|--|--|--|--|--|--|--||story|1|9|5|5|2|3|5|1||beach|8|8|4|5|3|5|9|2||vision|4|2|5|5|7|6|8|4|

本位置特征外，进一步引入了图表方向特征。如表9所示，与未引入方向特征的baseline模型相比，引入方向特征使得模型在准确率、召回率和F1分数上均提升了3%。

3.5.2 图表数据抽取

实验通过系统控制微调模块组合，深入探究视觉编码器（ViT）、语言解码器（LLM）及跨模态适配器（Adapter）在图表数据抽取任务中的差异化贡献。实验结果如表12所示。

表11 文本分类消融实验

Table 11 Text classification ablation

	precision	recall	F1-score
不包含方向特征	0.96	0.96	0.96
包含方向特征	**0.99**	**0.99**	**0.99**

注：加粗字体表示每列最优结果。

1）单一模块微调。ViT作为视觉编码器，其微调主要优化图表局部特征（如坐标轴、图例、数据点）的提取能力。如表12中实验1所示，与原始模型相比，对ViT微调后，模型的F1-score提升了28.13%。

LLM作为语言解码器，其微调显著提升模型对抽取结果的语义规范化能力（如数值单位转换、数据关系描述）。如表12中实验2所示，LLM微调后的precision（28.73%）高于ViT，说明其在减少文本生成错误（如数值偏差、格式混乱）方面更具优势，但过度依赖文本模板可能导致视觉细节丢失。

如表12中实验3所示，Adapter作为轻量级跨模态对齐模块，其微调虽然能缓解视觉—文本语义鸿沟，但由于缺乏对视觉或文本模块的深度优化，难以独立支撑复杂图表的细粒度特征对齐，性能提升有限。

2）双模块组合微调。如表12中实验4所示，通过ViT强化视觉特征提取，Adapter优化视觉—文本对齐，显著改善结构化图表中“视觉定位—语义映射”的准确性（如坐标轴标签与数值的对应关系）。然而，由于缺乏LLM的文本生成优化，最终输出的语义完整性仍受限（precision为30.62%，低于LLM + Adapter组合）。

如表12中实验5所示，通过LLM规范文本生成，Adapter增强跨模态对齐，可进一步提升图表数据转化能力。但LLM对视觉特征的间接依赖导致其在复杂图表中的视觉细节捕捉能力弱于ViT + Adapter组合（recall为42.88%，低于ViT + Adapter的47.91%）。

3）三模块联合微调。当同时微调ViT + Adapter + LLM时（可训练参数占比3.95%），模型实现“视觉特征提取—跨模态对齐—文本生成”的全流程优化，如表12中实验6所示，此时，precision、recall和F1-score均达到最优，分别为57.49%、63.65%和60.41%，验证了多模块联合微调对复杂图表数据抽取任务的必要性。

4 图表抽取系统设计与实现

为进一步落实本文的研究成果，本系统采用PyQt5进行图表数据抽取系统的开发，将所研究的基于规则的图表数据抽取方法、基于Qwen-VL微调的方法和现有的开源的图表数据抽取方法进行融

表12 LoRA微调消融实验结果

Table 12 Results of LoRA fine-tuning ablation experiments

/%

实验	微调部分	训练参数占比	precision	recall	F1-score
0	无	0	0.83	5.87	1.46
1	ViT	0.71	23.19	40.87	29.59
2	LLM	2.71	28.73	35.63	31.81
3	Adapter	0.53	17.91	33.50	23.34
4	ViT + Adapter	1.24	30.62	47.91	37.36
5	LLM + Adapter	3.24	38.79	42.88	40.73
6	ViT + Adapter + LLM	**3.95**	**57.49**	**63.65**	**60.41**

注：加粗字体表示每列最优结果。

合。如图 9 所示，整个系统界面分为两个大的 Tab 窗口：一个是基于规则的抽取方法，另一个是基于视觉语言大模型的抽取方法。

对于基于规则的数据抽取方法，按照本文方案先用 OCR 模型对图片中的文本进行提取和分类，再由目标检测模型识别出图例，并用图例将图片分割成不同的图表得到相应的数值后执行结构重建，因此需要对 OCR 模型与目标检测模型权重文件进行配置，直观展示图表文字识别结果及数据抽取结果。

图 9　系统界面—基于规则图表数据提取

Fig. 9　System interface-rule-based diagram data extraction

对于基于视觉语言大模型的图表数据抽取部分，系统参考了 Transformers 库中预训练模型的配置流程，除支持本文使用的 LoRA 微调的 Qwen-VL 模型外，还可以方便地嵌入其他预训练多模态大模型。其界面设计除了设置提示语外，还支持对其他一些配置参数的修改，例如温度和填充长度等。

针对上述两种方法，系统均在图表抽取结果的基础上，基于 Python 绘图库 Matplotlib 进一步设计了图表类型转换功能模块。该模块支持图表类型的转换以及参数的个性化设置。

5　结　论

本文的主要贡献在于：构建了中文图表数据集，为中文图表数据抽取研究提供了一份新的数据集支撑。同时，提出了两种基准模型：基于规则的图表数据抽取方法和基于大模型微调的数据抽取方法，为后续研究提供了参考标准。此外，还开发了图表抽取及类型转换系统，集成了多种图表提取模型，方便用户进行图表类型转换和数据抽取。

尽管取得了一定的成果，但当前数据集仍存在明显的局限性。本文提出的中文数据集目前仅限于条形图类别。随着图表理解任务向多类型（如折线图、饼图、散点图）和真实复杂场景发展，单一图表类型与人工合成数据可能限制模型的泛化能力。此外，基于大模型的模型在处理复杂图表结构和不规则文本时仍存在不足，将探索更鲁棒的模型结构和训练方法。最后，图表抽取及类型转换系统可进一步扩展功能，例如支持更多图表类型转换、图表风格修改等。未来工作中，计划引入真实场景图表及更多图表类型，以提升数据集的真实性和多样性。

综上，本文研究成果为中文图表数据抽取提供了新的解决方案，并为多模态数据分析领域的发展提供了有力支持。未来将继续探索更高效、更鲁棒的图表数据抽取方法，推动该领域的进一步发展。

参考文献（References）

Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, et al. 2023a.

Qwen technical report [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2309.16609.pdf

Bai J Z, Bai S, Yang S S, Wang S J, Tan S N, Wang P, et al. 2023b. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2308.12966.pdf

Bai S, Chen K Q, Liu X J, Wang J L, Ge W B, Song S B, et al. 2025. Qwen2.5-VL technical report [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2502.13923.pdf

Bajić F and Job J. 2022. Data extraction of circular-shaped and grid-like chart images. Journal of Imaging, 8(5): #136 [DOI: 10.3390/jimaging8050136]

Banerjee S and Lavie A. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, USA: Association for Computational Linguistics: 65-72

Ben Zaken E, Goldberg Y and Ravfogel S. 2022. BitFit: simple parameter-efficient fine-tuning for transformer-based masked language-models//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics: 1-9 [DOI: 10.18653/v1/2022.acl-short.1]

Carderas A, Yuan Y, Livnat I, Yanagihara R, Saul R, De Oca G M, et al. 2020. Automated data extraction of bar chart raster images [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2011.04137.pdf

Dehghani M, Mustafa B, Djolonga J, Heek J, Minderer M, Caron M, et al. 2023. Patch n' pack: NaViT, a vision transformer for any aspect ratio and resolution//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #106

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

Han Y C, Zhang C, Chen X, Yang X, Wang Z B, Yu G, et al. 2023. ChartLlama: a multimodal LLM for chart understanding and generation [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2311.16483.pdf

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He S, Ding L, Dong D Z, Zhang J and Tao D C. 2022. SparseAdapter: an easy approach for improving the parameter-efficiency of adapters//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics: 2184-2190 [DOI: 10.18653/v1/2022.findings-emnlp.160]

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, et al. 2019. Parameter-efficient transfer learning for NLP//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 2790-2799

Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, et al. 2022. LoRA: low-rank adaptation of large language models//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Huang K H, Zhou M Y, Chan H P, Fung Y, Wang Z H L, Zhang L Y, et al. 2024. Do LVLMs understand charts? Analyzing and correcting factual errors in chart captioning//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics: 730-749 [DOI: 10.18653/v1/2024.findings-acl.41]

Kafle K, Price B, Cohen S and Kanan C. 2018. DVQA: understanding data visualizations via question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5648-5656 [DOI: 10.1109/CVPR.2018.00592]

Kahou S E, Michalski V, Atkinson A, Kádár Á, Trischler A and Bengio Y. 2018. FigureQA: an annotated figure dataset for visual reasoning//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Kanroo M S, Kawoosa H S, Rana K and Goyal P. 2025. C3E: a framework for chart classification and content extraction. Computers and Electrical Engineering, 121: #109861 [DOI: 10.1016/j.compeleceng.2024.109861]

Lal J, Mitkari A, Bhosale M and Doermann D. 2023. LineFormer: line chart data extraction using instance segmentation//Proceedings of the 17th International Conference on Document Analysis and Recognition. San José, USA: Springer: 387-400 [DOI: 10.1007/978-3-031-41734-4_24]

Lester B, Al-Rfou R and Constant N. 2021. The power of scale for parameter-efficient prompt tuning//Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics: 3045-3059 [DOI: 10.18653/v1/2021.emnlp-main.243]

Li Y, Wu Q F, Liu J T and Zou J L. 2023. Leading weight-driven reposition relation network for figure question answering. Journal of Image and Graphics, 28(2): 510-521 (黎颖, 吴清锋, 刘佳桐, 邹嘉龙. 2023. 引导性权重驱动的图表问答重定位关系网络. 中国图象图形学报, 28(2): 510-521) [DOI: 10.11834/jig.211026]

Lialin V, Deshpande V, Yao X W and Rumshisky A. 2024. Scaling down to scale up: a guide to parameter-efficient fine-tuning [EB/

OL]. [2025-07-14]. http://arxiv.org/pdf/2303.15647.pdf

Liao M H, Wan Z Y, Yao C, Chen K and Bai X. 2020. Real-time scene text detection with differentiable binarization//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 11474-11481 [DOI: 10.1609/aaai.v34i07.6812]

Lin C Y. 2004. ROUGE: a package for automatic evaluation of summaries//Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics: 74-81

Liu F Y, Eisenschlos J, Piccinno F, Krichene S, Pang C X, Lee K, et al. 2023a. DePlot: one-shot visual language reasoning by plot-to-table translation//Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics: 10381-10399 [DOI: 10.18653/v1/2023.findings-acl.660]

Liu F Y, Piccinno F, Krichene S, Pang C X, Lee K, Joshi M, et al. 2023b. MatCha: enhancing visual language pretraining with math reasoning and chart derendering//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics: 12756-12770 [DOI: 10.18653/v1/2023.acl-long.714]

Liu H K, Tam D, Muqeeth M, Mohta J, Huang T H, Bansal M, et al. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #142

Ma W H, Zhang H S, Yan S, Yao G S, Huang Y C, Li H, et al. 2021. Towards an efficient framework for data extraction from chart images//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 583-597 [DOI: 10.1007/978-3-030-86549-8_37]

Masry A, Kavehzadeh P, Do X L, Hoque E and Joty S. 2023. UniChart: a universal vision-language pretrained model for chart comprehension and reasoning//Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore: Association for Computational Linguistics: 14662-14684 [DOI: 10.18653/v1/2023.emnlp-main.906]

Masry A, Long D X, Tan J Q, Joty S and Hoque E. 2022. ChartQA: a benchmark for question answering about charts with visual and logical reasoning//Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics: 2263-2279 [DOI: 10.18653/v1/2022.findings-acl.177]

Methani N, Ganguly P, Khapra M M and Kumar P. 2020. PlotQA: reasoning over scientific plots//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 1516-1525 [DOI: 10.1109/WACV45572.2020.9093523]

Obeid J and Hoque E. 2020. Chart-to-text: generating natural language descriptions for charts by adapting the transformer model//Proceedings of the 13th International Conference on Natural Language Generation. Dublin, Ireland: Association for Computational Linguistics: 138-147 [DOI: 10.18653/v1/2020.inlg-1.20]

Papineni K, Roukos S, Ward T and Zhu W J. 2002. BLEU: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics: 311-318 [DOI: 10.3115/1073083.1073135]

Pfeiffer J, Rücklé A, Poth C, Kamath A, Vulić I, Ruder S, et al. 2020. AdapterHub: a framework for adapting transformers//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. [s.l.]: Association for Computational Linguistics: 46-54 [DOI: 10.18653/v1/2020.emnlp-demos.7]

Sun K, Xiao B, Liu D and Wang J D. 2019. Deep high-resolution representation learning for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5686-5696 [DOI: 10.1109/CVPR.2019.00584]

Sung Y L, Nair V and Raffel C. 2021. Training neural networks with fixed sparse masks//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #1852

Wang P, Bai S, Tan S N, Wang S J, Fan Z H, Bai J Z, et al. 2024. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2409.12191.pdf

Wei J X, Xu N, Chang G Y, Luo Y, Yu B H and Guo R F. 2024. mChartQA: a universal benchmark for multimodal chart question answer based on vision-language alignment and reasoning [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2404.01548.pdf

Zhang L, Hu A W, Xu H Y, Yan M, Xu Y C, Jin Q, et al. 2024a. TinyChart: efficient chart understanding with visual token merging and program-of-thoughts learning [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2404.16635.pdf

Zhang L L, Huang M Y, Wang Q Y, Wang Y X, Wu W J, Liu J. 2024b. GoT-CQA: graph-of-thought guided compositional reasoning for chart question answering [EB/OL]. [2025-07-14]. http://arxiv.org/pdf/2409.02611.pdf

作者简介

马秋平，男，硕士研究生，主要研究方向为图表问答。
E-mail: maqiuping@stu.ppsuc.edu.cn
张琪，通信作者，女，副教授，主要研究方向为计算机视觉、模式识别。E-mail: qi.zhang@ppsuc.edu.cn
毕航烁，男，硕士研究生，主要研究方向为单图像超分辨率重建。E-mail: ppsucbhs@163.com
赵晓凡，女，副教授，主要研究方向为机器学习、算法理论。
E-mail: zhaoxiaofan@ppsuc.edu.cn