中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)09-3083-14

论文引用格式:Li Y Z, Fu L, Zhu L H, Luo Q D and Tu L. 2025. Multimodal large model-based method for generating visual Q&A data for electronic document images. Journal of Image and Graphics, 30(9):3083-3096(黎宇哲, 伏凌, 朱泠皞, 罗琪頔, 涂来. 2025. 多模态大模型面向电子文档视觉问答的数据生成. 中国图象图形学报, 30(9):3083-3096)[DOI:10.11834/jig.240610]

多模态大模型面向电子文档视觉问答的数据生成

黎宇哲[1],伏凌[1],朱泠皞[1],罗琪頔[1],涂来[2*]

1. 华中科技大学人工智能与自动化学院,武汉　430074; 2. 华中科技大学软件学院,武汉　430074

摘　要:**目的**　电子文档视觉问答数据生成技术旨在结合电子文档图像的文字内容与视觉信息,以生成问题及其对应答案。利用高质量的视觉指令微调数据集,可以显著提升多模态大型语言模型的文档阅读性能。目前,人工或模板方法生成的数据集存在数量不足和质量不高的问题。因此,本文设计了一种基于多模态大语言模型的电子文档图像问答数据生成方法。**方法**　提出一种基于多模态大型语言模型的大规模数据生成方法,包括4个关键步骤:自我提问与回答、数量与格式检查、数据过滤和一致性检验。通过输入电子文档图像及相应指令至多模态大型语言模型,初步生成多个问答对;进行数量与格式的检查;将合格的问答对及其对应图像和指令输入至多模态大型语言模型,以过滤掉与图像内容无关、回答错误或未使用外部知识的问答对;针对同一问答对,利用多模态大型语言模型生成多个不同表述的问题,并检查回答的一致性,以剔除回答不一致的问答对。**结果**　本文构建了一个高质量的数据集,包含324 546幅图像和2 036 263个问答对。通过对问答对正确率的随机抽样统计,结果显示正确率为91.34%。此外,还在DocVQA等文档类问答数据集上测试了该数据集对多模态大语言模型性能的提升作用。微调实验结果表明,在LLaVA-OV和Deepseek-VL模型上,基于本数据集的微调能够提升DocVQA数据集上的平均归一化编辑相似度,分别提高了1.4%和2.6%。消融实验进一步表明,去除数据过滤步骤后,模型性能下降了1.3%。通过与人工标注数据DocVQA的互补性实验,结果表明,在DocVQA训练集基础上加入部分视觉问答数据集进行训练后,模型性能比仅使用DocVQA训练集微调时提升了1.3%。此外,与现有方法生成的数据集进行性能对比时,本文方法生成的数据集在模型性能提升方面表现最为显著。后续的后处理实验也进一步证明了所提出的数据集在生成问答对时仍具有一定的提升空间。**结论**　本文提出的基于多模态大型语言模型的电子文档图像视觉问答数据生成方法,有效解决了现有数据集数量少、质量差的问题,显著提升了多模态大型语言模型的文档阅读理解能力。

关键词:多模态大模型;电子文档图像;视觉指令微调数据集;视觉感知理解;视觉文字

Multimodal large model-based method for generating visual Q&A data for electronic document images

Li Yuzhe[1], Fu Ling[1], Zhu Linghao[1], Luo Qidi[1], Tu Lai[2*]

1. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan　430074, China;

2. School of Software, Huazhong University of Science and Technology, Wuhan　430074, China

Abstract: **Objective** Recent advancements in multimodal large language models (MLLMs) have revolutionized the field of

收稿日期:2024-10-16;**修回日期**:2025-02-16;**预印本日期**:2025-02-23

*通信作者:涂来　tulai@hust.edu.cn

基金项目:国家重点研发计划课题(2022YFC2305103);国家自然科学基金项目(61202303);华中科技大学交叉研究支持计划项目(2023JCYJ047)

Supported by:National Key R&D Program of China(2022YFC2305103);National Natural Science Foundation of China(61202303)

visual question answering (VQA), especially in the domain of text-centric document understanding. These models have demonstrated remarkable improvements in tasks that involve the integration of visual and textual information, with several state-of-the-art models currently leading the field. One critical task in this area is the generation of VQA datasets for electronic documents, which entails combining the textual content embedded within document images with their visual components to generate meaningful and contextually relevant questions and corresponding answers. The integration of high-quality, fine-tuned datasets—specifically designed for multimodal instruction-following tasks—has greatly enhanced the document comprehension capabilities of MLLM. However, existing VQA datasets, which are typically generated using manual annotation techniques or templating methods, face significant challenges in scale and quality. These limitations impede the scalability and overall effectiveness of training datasets for multimodal models. Therefore, this paper proposes an innovative method to automatically generate image-based VQA datasets for electronic documents by utilizing an MLLM. The goal of this work is to address the existing gaps in dataset quality and quantity, thereby facilitating better training and fine-tuning of MLLM in the context of document-based visual question answering tasks. **Method** The proposed methodology involves the use of a large-scale data generation framework, powered by an MLLM. This framework is divided into four distinct stages: self-question generation, quantity and format verification, data filtering, and consistency validation. In the initial stage, the MLLM is tasked with generating multiple question-answer (Q&A) pairs by processing the input electronic document images alongside their corresponding textual descriptions. This stage capitalizes on the model's ability to simultaneously analyze both the visual and textual elements of the document, enabling it to generate a diverse array of questions that cover various aspects of the content, such as factual inquiries, inferential reasoning, and contextual understanding. The second stage focuses on ensuring that the generated Q&A pairs meet the required quantity and adhere to the correct formatting standards. This stage is critical for eliminating any inconsistencies, errors, or discrepancies in the formatting of the data, which could otherwise compromise the quality of the final dataset. In the third stage, data filtering is employed to refine the dataset by eliminating irrelevant or incorrect Q&A pairs. This process involves evaluating the generated Q&A pairs, along with their corresponding images and instructions, to identify and discard any irrelevant or improperly answered pairs. This step ensures that the dataset contains only high-quality questions that require multimodal reasoning capabilities for accurate responses. The final stage involves consistency validation, wherein the MLLM is used to generate multiple variations of the same Q&A pair. The objective of this stage is to verify that the answers remain consistent across different rephrasings of the same question. If inconsistencies in the answers are identified, then those pairs are discarded. This step not only ensures the reliability and accuracy of the dataset but also helps improve the robustness of the dataset by introducing diverse question formulations. By systematically applying these four stages, the proposed method enables the generation of a large-scale, high-quality VQA dataset for electronic documents, which can then be leveraged in fine-tune MLLMs to enhance their performance in document understanding tasks. **Result** In this study, a high-quality dataset was constructed, consisting of 324 546 images and 2 036 263 corresponding Q&A pairs. The overall correctness rate of 91.34% was achieved through random sampling of a sufficiently large number of images and their associated Q&A pairs, followed by manual verification of the selected samples. The effect of this dataset on improving the performance of MLLMs in document-based question answering tasks, such as DocVQA, was rigorously evaluated. Fine-tuning experiments on the LLaVA-OV and Deepseek-VL models demonstrated improvements of 1.4% and 2.6%, respectively, in average normalized Levenshtein similarity on DocVQA. In addition, ablation studies were conducted to assess the effectiveness of the data filtering process. These studies revealed that correctness filtering, relevance filtering, and external knowledge filtering each contributed to the enhancement of the performance of MLLMs when applied to the generated dataset. Interestingly, relevance filtering and external knowledge filtering did not conflict with one another. By contrast, applying both filtering methods simultaneously resulted in better model performance than when either one was applied alone. Furthermore, the entire data filtering process resulted in a 1.3% performance improvement for the Deepseek-VL model on the DocVQA dataset. Complementarity experiments with the DocVQA dataset showed a 1.3% performance gain when the model was fine-tuned on both the DocVQA dataset and a subset of the visual Q&A dataset. This finding demonstrated the ability of the generated dataset to complement manually labeled data and showcased the effectiveness of the synthetic data generation process. The superiority of the proposed dataset generation method was validated further through comparisons between the model performance

achieved using the ALLaVA and TG-Doc datasets and the model performance obtained using the generated data from the proposed method. Specifically, 1 million instruction samples were randomly selected from each of these datasets for full fine-tuning of the LLaVA-OV model. Experimental results indicated that the generated data from the proposed method led to the most significant improvement in model performance. Finally, while the proposed dataset resulted in some improvement in model performance, the overall gain was somewhat limited. A more in-depth analysis revealed that redundant characters in the generated answers——such as unnecessary phrasing——contributed to a degradation in performance. This issue was addressed by conducting a post-processing experiment using Qwen2. 5-14B. By removing redundant content from the model's outputs, the post-processing technique enhanced performance considerably, indicating that further refinement in the dataset generation process could yield even better results. **Conclusion** The proposed method for MLLM-driven generation of image-based VQA data for electronic documents effectively addresses the challenges of limited dataset size and poor data quality. The comprehensive evaluation of the dataset's effect on model performance, along with the successful implementation of data filtering and post-processing strategies, demonstrates the potential of this approach to improve the robustness and accuracy of multimodal models in document-based visual question answering tasks. Future work could further refine this process to eliminate redundant content and optimize the generated dataset for even better performance.

Key words: multimodal large model; electronic document image; visual instruction tuning dataset; visual perception and understanding; visual text

0 引　言

电子文档视觉问答数据生成是结合电子文档图像的文字内容与文档图像中的视觉信息，生成问题和对应答案对的技术。该技术通过微调多模态大语言模型(严昊 等,2023;刘成林 等,2023;兰红和张薄芬,2022)，提升模型在文档智能分析中的应用能力(刘禹良 等,2023)，具有重要的应用价值。

文档图像广泛应用于日常生活，结合文档图像和相关问题生成答案的融合多模态人工智能(artificial intelligence，AI)技术(兰红和张薄芬,2022;王峰等,2023)也日渐发展。提升此类模型性能的关键因素之一是数据量，因此获取大量高质量的电子文档视觉问答数据迫在眉睫。传统的文档问答数据生成方法主要依赖规则或模板驱动的方式，或者基于自然语言处理和光学字符识别(optical character recognition，OCR)技术的基本分析方法。例如，OCR-VQA数据集(Mishra 等,2019)便通过固定的问题模板生成问答对，尽管此方法能够大规模生成数据，但其多样性和准确性无法得到保证，生成的问答对通常无法与图像内容匹配。此外，这些方法往往需要人工干预，适用范围有限，且模板无法针对复杂格式和多种类型的文档图像提出合适的问题，表现较差。为了提高文档问答数据的生成质量，部分数据集采用了全程人工标注的方式构建，例如 DocVQA(Mathew 等,2021)通过人工标注生成数据集，尽管数据质量有保障，但由于人工标注的限制，无法大规模生成数据，且数据集中的文档图像和问答对数量十分有限。

多模态大型语言模型(multimodal large language model，MLLM)在以文本为核心的视觉问答(visual question answering，VQA)领域取得了显著进展(Ye 等,2023;Feng 等,2023a;刘禹良 等,2023;Liu 等,2024c;Long,2022 等)。这些研究通常尝试将视觉编码器与 MLLM 结合，通过中间模块如 Projector(Liu 等，2024b)、QFormer(Li 等，2023)和 Perceiver Resampler(Alayrac 等,2022)进行训练前对齐和指令微调，从而实现视觉语言理解。尽管这些方法在提升 MLLM 的视觉语言理解能力方面取得了一定成果，但仍面临不少问题，尤其是在以视觉为核心的数据生成流程中。为提高开源模型的性能，首先需要获取足够且高质量的训练数据。现有研究尝试通过增强模型的指令跟踪能力以弥补数据不足的问题。例如，mPLUG-DocOwl(Ye 等,2023)强调文档图像理解中对“文档结构”的理解，提出统一的结构学习方法，并创建了新的指令跟踪数据集;TextMonkey(Liu 等,2024d)采用移位窗口注意力机制过滤重要标记;DocPedia(Feng 等,2023a)和 HRVDA(Liu 等,2024a)则通过扩大输入分辨率来缩小 MLLM 与视觉文档理解之间的差距。然而，这些方法在数据生成方面仍然受限于小规模注释或单模态输入的局限，导致生成的数据面临规模小且质量差的问题。近期，针对

电子文档图像VQA数据不足的问题，已有研究提出一些解决思路，特别是使用MLLM作为数据生成器的尝试，如表1所示。Llama-GPT4（Peng等，2023）使用GPT-4生成指令跟踪数据进行微调；Synthetic Prompting（Shao等，2023）通过少量手工示例提示MLLM生成更多数据；TextSquare（Tang等，2024）则结合多模态理解能力的大模型建立了自动化数据生成流程，并在最终进行一致性检验来过滤数据。然而，这些方法在问答对的正确性、与图像的相关性以及是否涉及外部知识的过滤方面仍有不足。DocMatix（Laurençon等，2024）通过从PDFA（portable document format/A）中获取转录本，并使用Phi-3-small模型生成问答对，同时过滤幻觉内容（丢弃了15%的幻觉问答对），但缺乏一致性检验流程。这些方法在一定程度上扩展了数据集的规模，但由于生成的数据质量参差不齐，且规模较小，仍难以与闭源模型的数据质量相媲美。此外，针对通用VQA数据集生成的研究揭示了一些问题。现有的数据生成方法在图像字幕数据和VQA数据的生成粒度与范围上存在不一致。LLaVAR（Zhang等，2023）和TG-Doc（Wang等，2023b）通过整合OCR结果辅助生成更丰富的文本对话数据，但这些数据集的规模和精度不足以支持大规模模型的训练和微调。尽管Share-GPT4V（Chen等，2023）利用GPT-4构建了高质量的图像字幕数据集，但数据的粒度和范围仍有待提高；ALLaVA（Chen等，2024a）使用GPT-4从未标记的图像中生成推理指令和详细答案，但其数据规模依然较小。综上所述，当前的生成方法尚未能完全满足对高质量、多样化训练数据的需求。

为了解决上述问题，本文提出一种基于多模态大模型的大规模数据生成流程，旨在生成高质量的以文本为中心的视觉问答数据集。该流程包括4个主要步骤：自我提问与回答、数量与格式检查、数据过滤和一致性检验。首先，在自我提问与回答阶段，多模态大模型应用于分析图像内容并生成与图像文本相关的问题，同时给出简洁的回答。接着，在数量与格式检查阶段，系统会检查生成的问答对是否存在重复，并确保其格式符合要求，剔除不符合标准的问答对。随后，在数据过滤阶段，系统会筛选并剔除与图像内容无关或回答错误的问答对，同时排除那些未能有效利用大语言模型推理能力生成的问答对。最后，在一致性检验阶段，利用多模态大模型生成多种不同表述的问答对，并对其一致性进行检查，以提高数据的准确性和质量。

表1　多模态大语言模型生成问答对方法对比

Table 1　Comparison of methods for generating question-answer pairs using MLLM

方法	自动化流程	多模态理解能力	使用图像OCR信息	一致性检验	数据过滤机制
Llama-GPT4（Peng等，2023）	–	√	–	√（部分）	√（部分）
Synthetic Prompting（Shao等，2023）	√	–	–	–	–
LLaVAR（Zhang等，2023）	√	√	√	–	–
TG-Doc（Wang等，2023b）	√	√	√	–	–
ALLaVA（Chen等，2024a）	–	√	–	–	–
TextSquare（Tang等，2024）	√	√	–	√	–
DocMatix（Laurençon等，2024）	√	√	–	–	√
本文合成方法	√	√	√	√	√

注："√"表示有，"–"表示没有。

通过这一流程，能够有效减少模型在生成数据时可能出现的幻觉现象，增强数据的准确性与多样性，进而提升开源MLLM的性能。此外，从公共来源收集了大量富含文本的电子文档图像，并经过严格的去重与筛选，确保图像数据集的多样性和高质量。最终，本文构建了一个包含324 546幅图像和2 036 263个问答对的高质量数据集，为开源模型的指令微调提供了坚实的基础。

1　研究方法

提出的基于多模态大模型的电子文档视觉问答数据生成策略的核心流程如图1所示，主要包括4个步骤：自我提问与回答、数量与重复检查、数据过滤

和一致性检验。经过对比测试，最终选择了具有较强图像内容捕捉能力和优异指令跟踪能力的多模态大语言模型 InternVL1. 5-26B（Chen 等，2024b）作为数据生成的核心模型。

1. 1　自我提问与回答

基本流程如图 1 所示，首先提供电子文档文件图像。本文方法所使用的电子文档图像均来自 PubLayNet 数据集（Zhong 等，2019），该数据集包含扫描文档、数字文档、图像文档、表格文档以及电子邮件截图等类型的图像，所有图像均按照电子文档图像进行处理。提示库主要分为 3 部分：问答对生成提示库、数据过滤提示库和一致性检验提示库。每个提示库包含多种类型的提示，并且每条提示均经过测试验证其有效性。

对于每幅电子文档图像，从问答对生成提示库中随机选择一条提示，以生成问题和答案对。问答对生成提示库是一个指导多模态大语言模型根据电子文档图像生成符合要求问答对的指令集合。每条提示并不直接与具体图像内容相关，而是为多模态大模型提供生成目标的指导。例如："Play an image content analysis expert. First, analyze all the image contents in a comprehensive manner. Then, generate several meaningful and distinct factual, inferential, or open-ended questions with corresponding answers about the textual content of the image."，这些提示能够帮助模型聚焦于图像的不同维度，以确保生成的问题涵盖了图像的关键要素。随后，将图像和随机选取的提示输入多模态大语言模型，生成关于图像内容的多个问题和答案对，这些问题和答案不仅要符合语义逻辑，还要能够有效地反映图像的多样性和复杂性。

当前应用最广泛的多模态大语言模型结合了视觉和文本理解能力，能够处理和理解多种类型的数据，如文本和图像，并形成统一且连贯的表示。这些模型的核心在于其独特的架构设计，其中最为关键的两个部分是图像编码器和文本编码器。图像编码器负责从图像中提取独特的特征；文本编码器则专注于理解和生成自然语言。

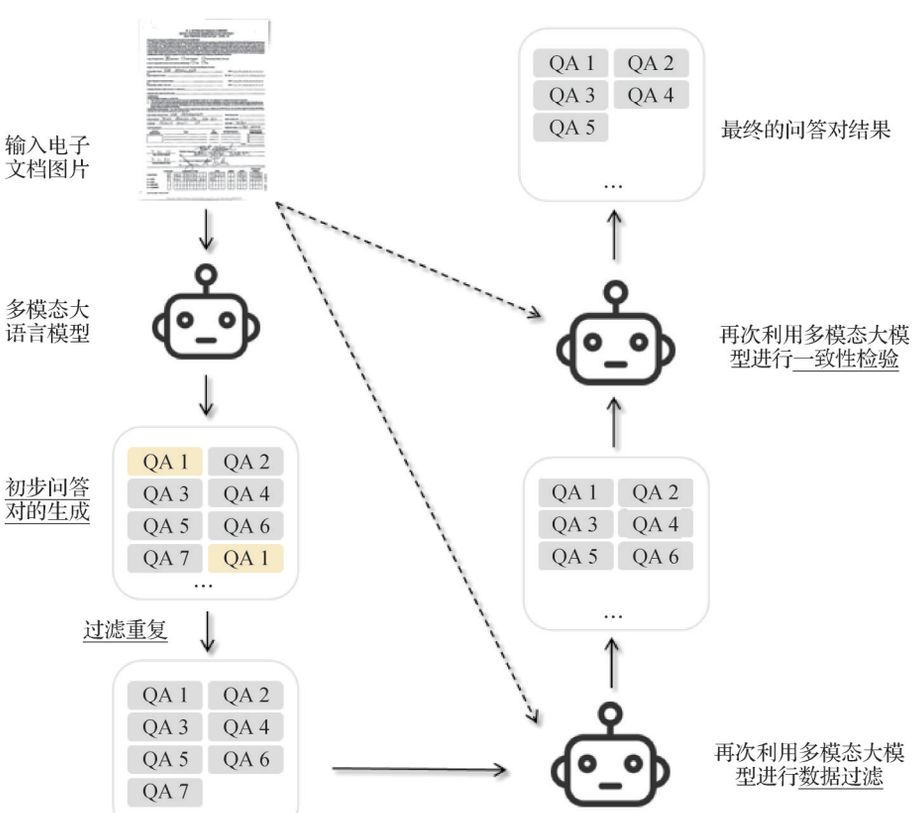


图 1　整体流程图

Fig. 1　Overall pipeline

提示库是多模态大语言模型的关键组成部分，涵盖了多种类型的提示。这些提示从不同的角度引导模型分析图像内容，生成多样化的问题并提供相应的答案。在为一幅图像生成问答对时，系统会从提示库中随机选择一条提示。这种随机性确保每次生成的问答对都是独立的，从而有效避免了重复性和预测性，提高了生成结果的多样性与灵活性。

1.2 数量与重复检查

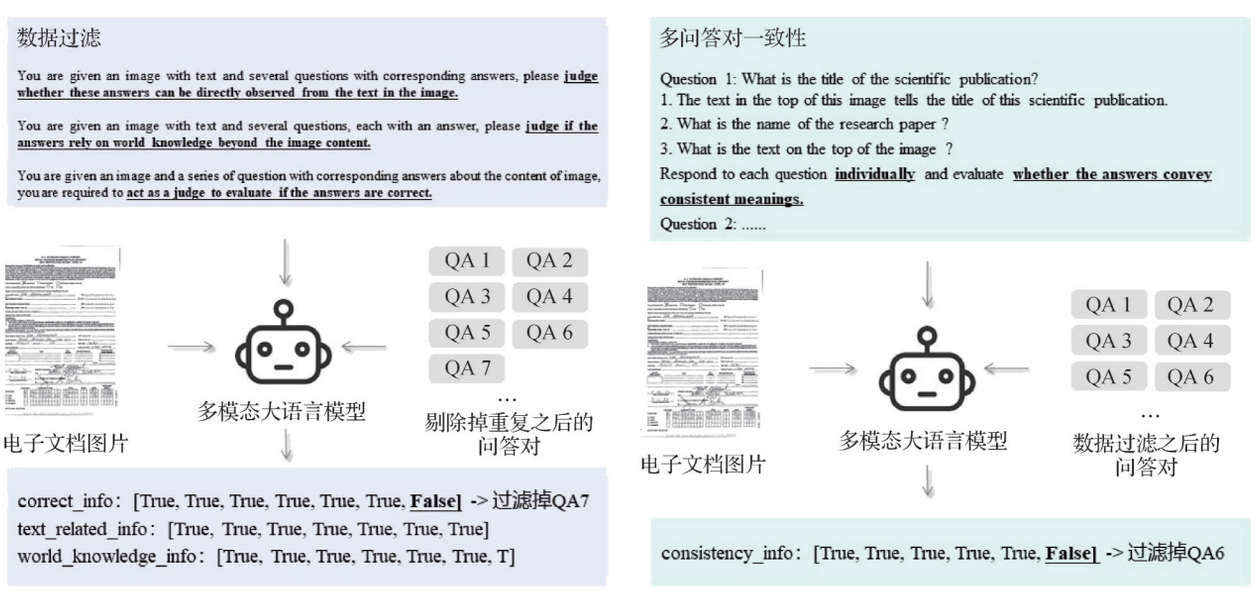
通过对第一步生成的问答对进行相似性分析，可以识别并删除重复内容，从而确保生成结果的唯一性。最直接的重复性检测方法是逐字比较，即通过字符串匹配技术识别完全相同的问题和答案对。如果两个问答对中的问题和答案在字符级别上完全一致，则判定为重复项。这种方法简单有效，尤其适用于生成过程中因随机性不足或模型输出多样性有限而导致的完全复制情况。一旦发现完全重复的问题和答案对，便可立即移除，确保结果集的纯净性。

然而，仅依赖字符级匹配并不足够，因为有些问题虽然表述不同，但实际上询问的是相同的信息，或者可以通过相同的答案来解答。这种情况需要进一步分析，因此，本文采用语义相似性评估来识别那些在意义层面重叠的问答对。通过利用多模态大语言模型的强大语义理解能力，可以对每幅图像对应的问答对进行语义分析，评估它们的相似度。如果两组问题和答案在语义层面非常接近，即多模态大语言模型认为它们实质上是重复的，那么，可以从每组相似的问答对中随机选择一个保留，其他则删除。这种随机选择有助于保持数据集的多样性和代表性，避免偏向任何特定的表述方式。

在删除重复问答对之后，为确保满足所需数量的独特问题和答案对，若有必要，系统会重复提示选择和生成过程，直到生成所需数量的独特问答对为止。

1.3 数据过滤

如图2(a)所示，数据过滤指剔除与图像内容无关的问答对、存在错误的问答对以及未使用大语言模型外部知识的问答对。该过程旨在确保生成的问答对不仅与图像内容紧密相关、回答准确无误，还能体现出多模态大语言模型内置的知识。数据过滤的目标是通过严格的筛选和验证，提升数据集的质量。



(a) 数据过滤流程　　(b) 一致性检验流程

图2 数据过滤与一致性检验流程图

Fig. 2 Data filtering and consistency check flow chart ((a) data filtering process; (b) consistency verification process)

1.3.1 相关性过滤

在创建图像相关的问答对时，首要任务是确保问题紧密围绕图像中的元素、情境或细节展开。例如，如果文档图像描述的是一项科学研究，包括研究课题、研究背景、研究目的以及研究方法等，那么问题应聚焦于研究的相关内容。然而，如果问题涉及图像的来源，或提及图像中未出现的研究员姓名，这类问题则与图像内容无关。筛选过程中应剔除这些

偏离主题的问题，确保每个问题都基于图像内容提出。

在经过第 1.2 节检查后的问答对与对应图像，将与第 1.1 节中提到的数据过滤提示库中的相关性过滤提示随机选取的提示（例如："You are provided with an image with textual information and several questions. Each question has a matched answer. Please judge whether these answers can be directly observed from the text in the image."）一起，送入多模态大语言模型进行判断。如果问题与图像内容不相关，则直接过滤。

1.3.2 外部知识过滤

多模态大语言模型的一大优势在于其能够整合广泛的背景知识，解决跨领域的问题。为了最大化这一优势，生成的问题和答案不仅应与图像内容密切相关，还需要合理运用外部知识。例如，对于一幅描述上世纪某位美国市民工作简历的图像，提出的问题可以是："图中人物从 1930 至 1935 年间的失业是否与美国第一次经济大萧条有关？"这一类问题和答案对不仅能够增强大模型对图像信息的理解能力，更重要的是，能够丰富大模型的外部知识库。

在数据过滤过程中，首先通过相关性过滤，确保生成的问题和答案与图像内容高度契合；然后，进一步进行外部知识验证。这一过程通过检查问题是否依赖于模型的外部知识来回答，确保答案不仅仅基于图像信息，还能够结合大语言模型的训练知识库。这样生成的数据能够极大地丰富大模型的世界知识。

经过相关性过滤后，所有符合条件的问答对将与相应图像以及第 1.1 节中提到的正确性过滤提示库中的相关性过滤提示（例如："You are given an image with text and several questions, each with an answer. Determine if the answers rely on world knowledge beyond the image content."）一同输入多模态大语言模型进行生成。如果问答对未能满足相关性或外部知识要求，则将其过滤掉，确保最终生成的问答对在内容相关性和外部知识应用上均达到高标准。

1.3.3 正确性过滤

数据过滤过程的最后一步是对问答对进行正确性验证，剔除错误的问答对，以确保最终结果的准确性。错误的回答可能源于模型对问题的误解，或虽有正确答案，但未能准确回应提问的实际意图。例如，若问题询问"图中表格的含义？"而回答却是表格的名称，尽管该答案无误，但与表格的实际内容无关。此时，虽然问题与图像有关，回答却未能紧扣问题的核心，因此需要被标记并修正。

为了进一步提升问答对的整体质量，本文将在此基础上再次利用多模态大语言模型检查每一对问答，若发现错误的回答，则直接过滤掉。经过相关性过滤和外部知识过滤后的问答对及其对应图像，将与第 1.1 节提到的正确性过滤提示库中随机选取的提示（例如："You are given an image and a series of questions with corresponding answers about the content of the image. You are required to act as a judge to evaluate if the answers are correct."）一同送入多模态大语言模型进行判定，若发现不合格的问答对，则予以过滤。

1.4 一致性检验

对过滤后的问答对进行一致性检查，旨在进一步验证回答的正确性。在 1.3.3 节中，本文发现仅通过正确性过滤无法彻底剔除错误的问答对，因为对于多模态大语言模型而言，许多实际错误的问答对在传统的正确性过滤过程中并未被识别和剔除。一致性检验首先将过滤后的问答对中的问题与对应图像，以及从 1.1 节中提到的一致性检验提示库中随机选取的一条提示（例如："You are given an image and a corresponding question about the content of the image. You are required to generate several questions with the same meaning but different wording."）一同送入多模态大语言模型，生成多个表达方式不同但答案相同的问题。

例如，给定一幅科学出版物的封面及其对应问题："What is the text at the top of this image that indicates the title of the scientific publication?"生成的变体问题可能包括："What is the title of the scientific paper at the top of this image?"、"What is the name of the research paper?"、"What text at the top of the image indicates the title?"等，所有这些问题的答案都应该是该科学出版物的标题。接着，通过多模态大语言模型对这些问题的答案进行检验，如果所有答案都正确，则保留该问答对，否则将其剔除。

该方法通过重新构建原始问题，生成多种语义上等效的问法，进而确保信息的准确性与逻辑性，特别是在处理可能存在偏差的数据时。

2 数据集介绍

如表2所示，本文提出的数据集共包含324 546幅电子文档图像，均来自PubLayNet数据集（Zhong等，2019）。PubLayNet是文档图像版面分析的大型数据集，其布局通过多边形边框进行分割标注。与这些图像对应的问答对共有2 036 263对，平均每幅电子文档图像包含6.27个问答对。与之前的OCR-VQA（Mishra等，2019）使用的模板生成问答数据集相比，本文数据集在多样性上有明显进步。与人工标注的数据集如DocVQA（Mathew等，2021）、InfographicVQA（Mathew等，2022）和VisualMRC（Tanaka等，2021）相比，本文数据集在图像和问答对的数量上也具有明显优势，尤其是平均每幅图像包含的问答对数。虽然WebSRC（Chen等，2021）每幅图像包含的问答对数高达62个，但一幅图像所需的有效信息不需要如此多的问答对来描述，这意味着会存在许多琐碎且无意义的问题。而与用大模型生成的UniDoc（Feng等，2023b）数据集相比，后者每幅图像仅包含一个问答对，无法有效提取图像中的信息。尽管本文数据集的总体规模相较于用多模态大模型大规模生成的TextSquare（Tang等，2024）和DocMatix数据集较小，但每幅图像的平均问答对数接近于其两倍，显示出更高的图像信息提取能力。

表2 电子文档问答数据集统计

Table 2 Statistics of electronic document question answering datasets

数据集	图像数量/幅	指令数量	平均每幅图像包含指令数	指令生成方式
OCR-VQA（Mishra等，2019）	207 572	1 002 146	4.82	模板
DocVQA（Mathew等，2021）	12 767	50 000	3.91	人工
InfographicVQA Mishra等，2019）	5 485	30 035	4.63	人工
VisualMRC（Tanaka等，2021）	10 197	30 562	3.00	人工
WebSRC（Chen等，2021）	6 447	400 498	62.12	人工
ALLaVA（Chen等，2024a）	661 573	1 425 097	2.15	大模型生成
UniDoc(Fine-tuning)（Feng等，2023b）	150 000	150 000	1.00	大模型生成
TextSquare(PDF)（Tang等，2024）	1 000 000	3 500 000	3.50	大模型生成
DocMatix	2 444 750	9 500 000	3.88	大模型生成
本文数据集	324 546	2 036 263	6.27	大模型生成

对本文所提数据集随机抽取50幅图像对应302个问答对进行人工正确率统计，结果为91.34%，证明了利用本文方法所构建数据集的高质量。本文提出的数据集中问题的单词数（token）分布如图3所示，可以看到基本呈正态分布，单词数在8～12区间内的图像数量占大部分，而回答的单词数分布如图4所示。由于本文要求回答尽量简短，因此单词数小于6时，对应的图像数量较多。可以证明，本文的提示对多模态大语言模型的表现具有明显的规范作用。当问题无法以很少的单词数回答，回答的单词数超过6时，每个单词数对应的图像数量再次呈现正态分布，回答的平均单词数为22个。

本文抽样统计了近1 000个问答对，并人工按照以下5个类别进行分类：

1）结构化信息提取问题。指的是问题集中在提取某些特定的、结构化的信息（如文章标题、作者、出版日期等），通常回答是标准化的文本或数字，如“What is the title of the journal?”。

2）数字和图表理解问题。指的是问题集中在对图表、表格等视觉化信息的解析上，通常要求提取数字、标题或缩略语的含义如“What is the main graph in the image titled?”。

3）方法和过程问题。指的是问题涉及到实验或研究过程的描述，通常需要解释某个方法或步骤的目的、过程或结果，如“What is the purpose of RT-

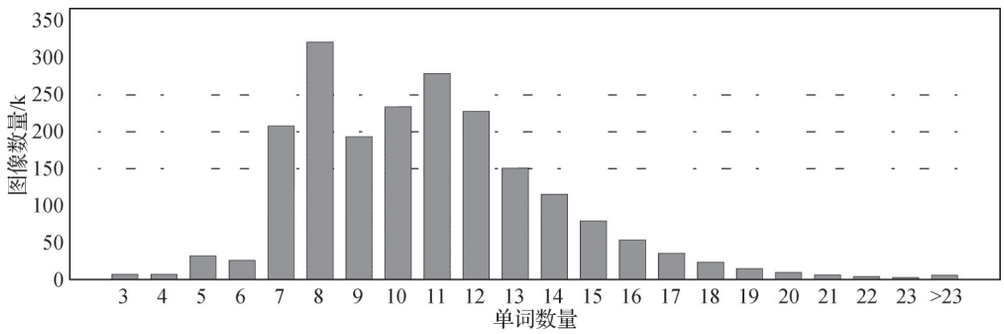


图3　问题的单词数分布

Fig. 3　Words distribution of questions

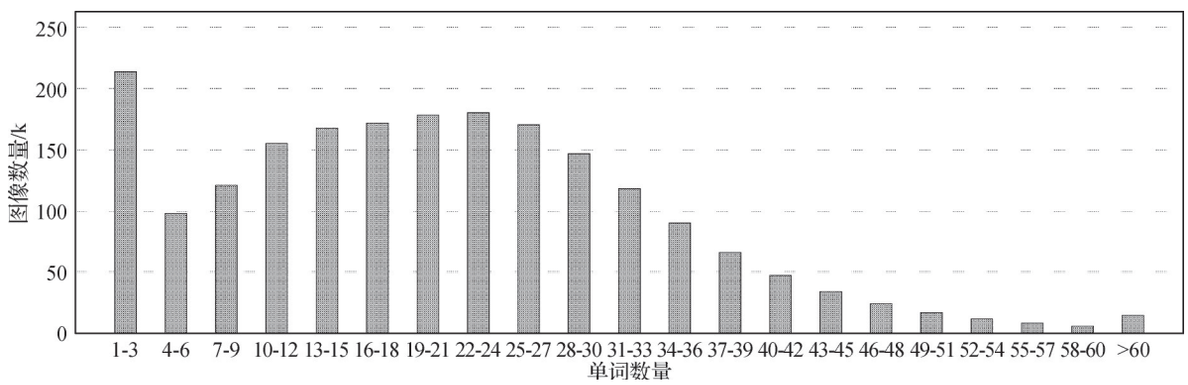


图4　回答的单词数分布

Fig. 4　Words distribution of answers

PCR analysis in this context?”。

4)概念与重要性问题。指的是问题着重探讨研究或数据的意义,通常要求提供解释或推论,如“What is the significance of FNAB in defining lesions without characteristic imaging appearance?”。

5)文章焦点和主题问题。指的是问题关注文献的主要内容和研究主题,要求总结或提取核心焦点,如“What is the main focus of the article?”。以上5个类别的分布如图5所示。

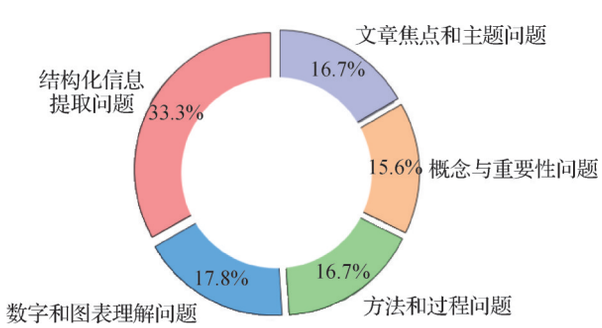


图5　问题类别抽样统计

Fig. 5　Sampling statistics of problem categories

3　实验设置

本文选取多个通用多模态大语言模型,包括轻量化LLaVA-OV-0.5B模型(Li等,2024a)与deepseek-VL模型(Lu等,2024),使用本文数据集分别进行了全量微调与LoRA(low-rank adaptation)微调(Hu等,2022)。为了评估微调后的模型在处理文档视觉问答任务中的表现,本文选取了与文档相关的数据集DocVQA(Mathew等,2021)、InfographicVQA(Mathew等,2022)和ChartQA(Masry等,2022),具体介绍见3.1节。

3.1　数据集

DocVQA是一个专注于文档图像的视觉问答数据集。它由50 000个问题和12 000多幅文档图像组成,旨在通过文档内容提取和使用来回答由人类定义的高级任务。DocVQA用于研究机器如何“理解”文档图像并回答有关它们的问题。

ChartQA是一个大规模基准数据集,它专注于图表的视觉和逻辑推理问答。该数据集包含9 600个

人工编写的问题和23 100个由人工编写的图表摘要生成的问题。ChartQA的特点是它不仅包含了真实世界的图表和人工编写的问答对，而且还包括了详细的标注，如边界框和数据点的具体信息。这要求模型不仅要理解文本，还要能解析复杂的视觉元素，并结合逻辑进行推理。ChartQA的提出是为了解决现有数据集在复杂推理问题上的不足，这些问题通常是基于模板的，答案来自固定词汇。ChartQA的数据集结构是为了应对这些挑战而设计的。

InfographicVQA是一个专注于图表信息的理解和推理的视觉问答数据集。这个数据集包含了多样的图表信息及相应的自然语言问题和答案注释。数据集中的问题需要方法联合推理文档布局、文本内容、图形元素和数据可视化。InfographicVQA数据集强调需要基本推理和基本算术技能的问题。该数据集包含30 035个问答对，分布在5 485幅图像上，训练集中有23 946个问答对分布在4 406幅图像上，验证集中有2 801个问答对分布在500幅图像上，测试集中有3 288和问答对分布在579幅图像上。这些问题包括基于表格、图形和可视化的问题，以及需要结合多个线索的问题。

3.2 定量评价

本文所有在ChartQA数据集上的测试指标为宽松的准确度(relaxed accuracy)，在DocVQA和InfographicVQA数据集上的测试指标均为平均归一化Levenshtein相似度(average normalization Levenshtein similarity，ANLS)。

4 实验分析

首先选择轻量化预训练大模型LLaVA-OV(Li等，2023)，选择了参数量仅有0.5 B(5亿)的LLaVA-OV-0.5，因此使用本文合成数据集进行全量微调，分别在使用30万条与100万条指令数据量的情况下进行测试，结果如表3所示。在使用30万条指令进行微调之后，LLaVA-OV-0.5B在DocVQA测试集上有了0.5%的提升，同时在InfographicVQA和ChartQA类文档数据集的测试集上也有小幅提升；当模型微调使用的数据量上升到100万之后，模型性能有了明显的提升：在DocVQA测试集上有1.4%的提升，在InfographicVQA和ChartQA类文档数据集的测试集上也有超过1%的提升。由此可见，本文构造的数据集对于多模态大模型文档阅读能力的有效提升作用。

其次，使用了Deepseek-VL(Lu等，2024)模型，具体模型是总参数量为1.5 B(15亿)的Deepseek-VL-1_3b-chat，在本文的数据集上进行LoRA(low-rank adaptation)微调，可训练参数量为0.38%，分别在训练数据量为100万与所有指令数据的情况下进行了测试，结果如表4所示。由于该模型原来的文档阅读能力较弱，因此本文所构造数据集对其性能提升效果更为明显。首先在训练数据量为100万时，模型在DocVQA上有了2.1%的提升，在InfographicVQA与ChartQA上也有1%左右的提升；当数据量为本文所构造的全部指令时，模型在DocVQA上的表现提升了2.6%，在InfographicVQA和ChartQA上分别有1.4%和1.3%的提升。

表3 LLaVA-OV-0.5B模型微调实验结果

Table 3 Experimental results of LLaVA-OV-0.5B model fine-tuning

/%

模型/训练数据量	DocVQA	InfographicVQA	ChartQA
LLaVA-OV-0.5B	57.1	32.1	55.8
LLaVA-OV-0.5B/30w	57.6	32.5	56.1
LLaVA-OV-0.5B/100w	58.5	33.4	57.0

注：30w：使用30万条指令；100w：使用100万条指令。

表4 Deepseek-VL-1_3b-chat模型微调实验结果

Table 4 Experimental results of deepseek-vl-1_3b-chat model fine-tuning

/%

模型/训练数据量	DocVQA	InfographicVQA	ChartQA
Deepseek-VL-1_3b-chat	26.5	15.8	38.9
Deepseek-VL-1_3b-chat/100w	28.6	16.8	39.8
Deepseek-VL-1_3b-chat/all	29.1	17.2	40.2

注：100w：使用100万条指令；all：使用全部指令。

4.1 消融实验

本文进行了消融实验，以验证数据过滤过程中各个步骤的有效性，并重点评估每一项过滤操作对生成数据集质量的影响。为了量化每个过滤操作的效果，通过逐步移除每个过滤步骤，评估其对数据集质量的潜在影响。为确保对比的一致性，选取了表5

中100万条指令数据训练的结果作为对照组。基于此，分别生成了4种指令数均为100万的数据集：1)完全不进行数据过滤的数据集(表5第1行)；2)仅进行单一数据过滤操作的数据集(表5第2–4行)；3)仅进行相关性过滤与外部知识过滤的数据集(表5第5行)；4)应用所有过滤操作的数据集(表5第6行)。

随后，使用Deepseek-VL-1_3b-chat模型，在与对照组相同的训练设置下，对新生成的数据集进行了训练。实验结果如表5所示，经过两次操作后的数据集在提升模型视觉问答性能方面均表现出积极作用，而整个数据过滤过程使得DocVQA的性能提升了1.3%。另外，同时使用相关性过滤与外部知识过滤对模型性能的提升作用比单独使用其中一项时明显更好，验证了本文所提出的相关性过滤与外部知识过滤均对最终生成数据质量有提升作用。

表5　数据过滤流程的消融实验

Table 5　Ablation experiments on data filtering process

正确性过滤	相关性过滤	外部知识过滤	DocVQA/%
–	–	–	27.3
√	–	–	28 3
–	√	–	27 9
–	–	√	27 5
–	√	√	28.1
√	√	√	28.6

注："√"代表使用相应策略，"–"代表不使用。

4.2　与人工标注数据的互补性实验

为了验证提出的数据集在视觉问答任务中的有效性，对本文生成的视觉问答数据集与现有的人工标注数据集DocVQA进行对比实验分析。选取了一种多模态大型语言模型LLaVA-1.5(Liu等，2024b)，该模型在预训练阶段并未包含文档理解相关的数据集。实验设计包括两个阶段：首先，使用DocVQA数据集的训练集对模型进行独立训练；其次，将DocVQA训练集与从本文提出的数据集中随机选取的问答对进行合并，并进行LoRA微调。实验结果显示，模型在DocVQA测试集上的零样本推理性能为16.6%，表明模型在文档理解方面存在明显的性能不足。经过仅使用DocVQA训练集训练后，模型性能提升至30.4%。进一步地，构建了一个混合数据集，具体来说，在DocVQA数据集中加入了本文构建的数据集，但是每幅图像仅取一个问答对最为新的训练集，因此这个数据集上的图像数量为337 313幅(来自DocVQA的12 767幅与来自本文数据集的324 546幅)，指令数为374 546条(来自DocVQA的50 000条与来自本文数据集的324 546条)。结果如表6所示，用上述数据集对模型进行LoRA微调后，性能进一步提升至31.7%。这一结果表明了本文提出的数据集与人工标注数据集之间具有显著的互补性，能够显著提升模型在视觉问答任务中的性能。

表6　与人工标注数据的互补性实验

Table 6　Complementarity test with manually labeled data

模型/训练数据量	DocVQA /%
LLaVA_5-7b-instruct/zero-shot	16.6
LLaVA_5-7b-instruct/DocVQA	30.4
LLaVA_5-7b-instruct/DocVQA + 本文数据集(part)	31.7

4.3　与现有生成方法所生成数据集的性能对比

为了验证本文生成方法的优越性，本文选择与表1中的ALLaVA和TG-Doc数据集进行模型性能对比。具体而言，采用与表3中相同的LLaVA-OV-0.5模型，并为了确保对比的公平性，分别从上述数据集中随机选择100万条指令数据进行全量微调，并与本文生成的数据集进行对比。相关结果如表7所示。可见，本文方法所生成的数据集具有优越性。

表7　与现有方法所生成数据集质量比较实验

Table 7　Experimental comparison of the quality of datasets generated by existing methods

数据集/100 W	DocVQA/%
ALLaVA	58.3
TG-Doc	57.9
本文数据集	58.5

4.4　后处理实验

在本文的实验中，虽然本文提出的数据集对模型性能有一定的提升效果，但总体提升并不显著。通过进一步分析，发现生成结果中的冗余字符可能是影响模型表现的潜在原因。例如，问题为"What is the page number of the title 'Enrollment' in CONTENTS?"，目标答案应为数字"8"，然而生成的

答案却是“The page number of the title ‘Enrollment’ in CONTENTS is 8”。这类冗长的答案虽然在语义上是正确的，但多余的字符增加了模型处理的复杂性，导致评价指标下降。

为了进一步验证这一假设，本文进行了后处理实验。分别选择了在表3与表4中用100万指令微调后的LLaVA-OV-0.5B与Deepseek-VL-1_3b-chat模型在DocVQA测试集上测试的结果。随后，本文使用Qwen2.5-14B（Hui等，2024）对模型的输出进行后处理。具体而言，将问题与对应的答案输入Qwen2.5-14B，并判断结果中是否存在冗余内容。如果答案中包含冗余，Qwen2.5-14B会提取有效信息并返回一个更加简洁的结果；如果答案没有冗余，则跳过处理。

后处理实验结果如表8所示。通过对比后处理前后的实验结果，可以看到，后处理能够有效减少冗余内容，并显著提升模型在测试集上的表现。这一结果表明，本文所提出的数据生成方法在生成问答对时仍存在一定的提升空间。

表8　后处理实验

Table 8　Post-processing experiment

模型/训练数据量	DocVQA/%
LLaVA-OV-0.5B	57.1
LLaVA-OV-0.5B/100w	58.5
LLaVA-OV-0.5B/post-process	61.4
Deepseek-VL-1_3b-chat	26.5
Deepseek-VL-1_3b-chat/100 w	28.6
Deepseek-VL-1_3b-chat/post-process	32.3

5　结　论

在多模态大型语言模型应用于文档图像生成问答对任务中，本文指出了现有方法的不足，包括数据集数量不足和质量不理想。为解决这些问题，本文提出一种自动化方法，利用多模态大型语言模型生成文档图像的问答对，相较于传统人工方法，显著扩大了数据集规模。以往人工生成的数据集通常包含数千幅图像，而DocVQA数据集包含12 767幅图像，本文生成的数据集则超过30万幅。与基于模板生成的问答对相比，后者问题类型单一，无法适应每幅文档图像的独特内容。本文方法通过多模态大型语言模型对齐不同模态信息的能力，深入挖掘图像中的文字内容，生成多样化的问题。

为提高问答对质量，本文引入了数据过滤与一致性检验流程，包含相关性过滤、外部知识过滤和正确性过滤。通过消融实验，证明了这些步骤在剔除低质量问答对方面的有效性，并确保生成数据对模型性能的提升。进一步对比表3和表4的结果显示，在相同设置下，本文方法在多模态大型语言模型文档阅读能力较弱时，性能提升更为显著。最终对比实验验证了本文数据集的优越性，且微调后，模型对文档任务的处理能力得到了提升。如图6所示，对于在训练前多模态大模型无法正确回答的问题，经过本文数据集微调后，模型能够给出正确的答案。

然而，本文也注意到数据集对多模态大型语言模型性能提升的局限性。一方面，数据量仍有待增加，后续研究可通过引入更多文档图像数据来扩充

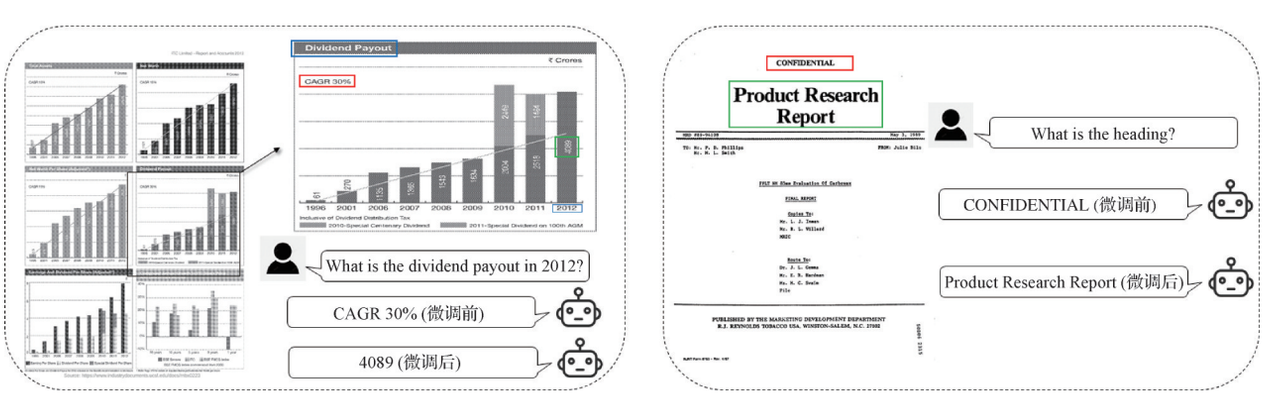


图6　本文提出数据集对多模态大模型在针对文档任务上性能优化案例

Fig. 6　This paper proposes a case study on the performance optimization of multimodal large models on document tasks

数据集。本文数据集的平均回答单词数为21个，而DocVQA等数据集平均为2.35个，可见本文数据集与DocVQA存在较大的域差异（domain gap）。这限制了本文数据集在大幅提升模型在DocVQA等数据集上性能的潜力。

参考文献（References）

Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T D, Gong Z T, Samangooei S, Monteiro M, Menick J L, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Bińkowski M, Barreira R, Vinyals O, Zisserman A and Simonyan K. 2022. Flamingo: a visual language model for few-shot learning//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1723

Chen G H, Chen S N, Zhang R F, Chen J Y, Wu X B, Zhang Z Y, Chen Z H, Li J Q, Wan X and Wang B Y. 2024a. ALLaVA: harnessing GPT4V-synthesized data for lite vision-language model [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2402.11684.pdf

Chen L, Li J S, Dong X Y, Zhang P, He C H, Wang J Q, Zhao F and Lin D H. 2023. ShareGPT4V: improving large multi-modal models with better captions//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer [DOI: 10.1007/978-3-031-72643-9_22]

Chen X Y, Zhao Z H, Chen L, Ji J B, Zhang D Y, Luo A, Xiong Y X and Yu K. 2021. WebSRC: a dataset for web-based structural reading comprehension//Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics: 4173-4185 [DOI: 10.18653/V1/2021.EMNLP-MAIN.343]

Chen Z, Wu J N, Wang W H, Su W J, Chen G and Xing S. 2024b. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24185-24198 [DOI: 10.1109/CVPR52733.2024.02283]

Feng H, Liu Q, Liu H, Tang J Q, Zhou W G, Li H Q and Huang C. 2023a. DocPedia: unleashing the power of large multimodal model in the frequency domain for versatile document understanding. Science China Information Sciences, 67(12): #220106 [DOI: 10.1007/s11432-024-4250-y]

Feng H, Wang Z J, Tang J Q, Lu J H, Zhou W G, Li H Q and Huang C. 2023b. UniDoc: a universal large multimodal model for simultaneous text detection, recognition, spotting and understanding [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2308.11592.pdf

Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L and Chen W. 2022. LoRA: Low-rank adaptation of large language models//Proceedings of the International Conference on Learning Representations (ICLR 2022). Online: OpenReview

Hui B Y, Yang J, Cui Z Y, Yang J X, Liu D Y H, Zhang L, Liu T Y, Zhang J J, Yu B W, Lu K M, Dang K, Fan Y, Zhang Y C, Yang A, Men R, Huang F, Zheng B, Miao Y B, Quan S H R, Feng Y L, Ren X Z, Ren X C, Zhou J R and Lin J Y. 2024. Qwen2.5-coder technical report [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2409.12186.pdf

Lan H and Zhang P F. 2022. Question-guided spatial relation graph reasoning model for visual question answering. Journal of Image and Graphics, 27(7): 2274-2286 (兰红, 张蒲芬. 2022. 问题引导的空间关系图推理视觉问答模型. 中国图象图形学报, 27(7): 2274-2286) [DOI: 10.11834/jig.200611]

Laurençon H, Marafioti A, Sanh V and Tronchon L. 2024. Building and better understanding vision-language models: insights and future directions [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2408.12637.pdf

Li B, Zhang Y H, Guo D, Zhang R R, Li F, Zhang H, Zhang K C, Zhang P Y, Li Y W, Liu Z W and Li C Y. 2024a. LLaVA-oneVision: easy visual task transfer [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2408.03326.pdf

Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #814 [DOI: 10.5555/3618408.3619222]

Liu C H, Yin K, Cao H Y, Jiang X H, Li X, Liu Y S, Jiang D Q, Sun X and Xu L L. 2024a. HRVDA: high-resolution visual document assistant//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 15534-15545 [DOI: 10.1109/CVPR52733.2024.01471]

Liu C L, Jin L W, Bai X, Li X H and Yin F. 2023. Frontiers of intelligent document analysis and recognition: review and prospects. Journal of Image and Graphics, 28(8): 2223-2252 (刘成林, 金连文, 白翔, 李晓辉, 殷飞. 2023. 文档智能分析与识别前沿：回顾与展望. 中国图象图形学报, 28(8): 2223-2252) [DOI: 10.11834/jig.221112]

Liu H T, Li C Y, Li Y H and Lee Y J. 2024b. Improved baselines with visual instruction tuning//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26286-26296 [DOI: 10.1109/CVPR52733.2024.02484]

Liu Y L, Li H L, Bai X and Jin L W. 2023. A brief analysis of ChatGPT: historical evolution, current applications, and future prospects. Journal of Image and Graphics, 28(4): 893-902 (刘禹良, 李鸿亮, 白翔, 金连文. 2023. 浅析ChatGPT：历史沿革、应用现状及前景展望. 中国图象图形学报, 28(4): 893-902) [DOI: 10.11834/jig.230110]

Liu Y L, Li Z, Huang M X, Yang B, Yu W W, Li C Y, Yin X C, Liu C L, Jin L W and Bai X. 2024c. OCRBench: on the hidden mystery of OCR in large multimodal models. Science China Information

Sciences, 67(12): #220102 [DOI: 10.1007/s11432-024-4235-6]

Liu Y L, Yang B, Liu Q, Li Z, Ma Z Y, Zhang S and Bai X. 2024d. TextMonkey: an OCR-free large multimodal model for understanding document [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2403.04473.pdf

Long S B, Qin S Y, Panteleev D, Bissacco A, Fujii Y and Raptis M. 2022. Towards end-to-end unified scene text detection and layout analysis//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 1039-1049 [DOI: 10.1109/CVPR52688.2022.00112]

Lu H Y, Liu W, Zhang B, Wang B X, Dong K, Liu B, Sun J X, Ren T Z, Li Z S, Yang H, Sun Y F, Deng C Q, Xu H W, Xie Z D and Ruan C. 2024. DeepSeek-VL: towards real-world vision-language understanding [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2403.05525.pdf

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. Dublin, Ireland: ACL: 2263-2279 [DOI: 10.18653/V1/2022.FINDINGS-ACL.177]

Mathew M, Bagal V, Tito R, Karatzas D, Valveny E and Jawahar C V. 2022. InfographicVQA//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2582-2591 [DOI: 10.1109/WACV51458.2022.00264]

Mathew M, Karatzas D and Jawahar C V. 2021. DocVQA: a dataset for VQA on document images//Proceedings of 2021 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2199-2208 [DOI: 10.1109/WACV48630.2021.00225]

Mishra A, Shekhar S, Singh A K and Chakraborty A. 2019. OCR-VQA: visual question answering by reading text in images//Proceedings of 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney, Australia: IEEE: 947-952 [DOI: 10.1109/ICDAR.2019.00156]

Peng B L, Li C Y, He P C, Galley M and Gao J F. 2023. Instruction tuning with GPT-4 [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2304.03277.pdf

Shao Z H, Gong Y Y, Shen Y L, Huang M L, Duan N and Chen W Z. 2023. Synthetic prompting: generating chain-of-thought demonstrations for large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #1273 [DOI: 10.5555/3618408.3619681]

Tanaka R, Nishida K and Yoshida S. 2021. VisualMRC: machine reading comprehension on document images//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI: 13878-13888 [DOI: 10.1609/AAAI.V35I15.17635]

Tang J Q, Lin C H, Zhao Z, Wei S, Wu B H, Liu Q, Feng H, Li Y, Wang S Q, Liao L, Shi W, Liu Y L, Liu H, Xie Y, Bai X and Huang C. 2024. TextSquare: scaling up text-centric visual instruction tuning [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2404.12803.pdf

Wang F, Shi F Y, Zhao J, Zhang X S and Wang X F. 2023. Answer mask-fused visual question answering model. Journal of Image and Graphics, 28(11): 3562-3574 (王峰, 石方宇, 赵佳, 张雪松, 王雪枫. 2023. 融合答案掩码的视觉问答模型. 中国图象图形学报, 28(11): 3562-3574) [DOI: 10.11834/jig.211137]

Wang Y H, Zhou W G, Feng H, Zhou K Y and Li H Q. 2023b. Towards improving document understanding: an exploration on text-grounding via MLLMs [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2311.13194.pdf

Yan H, Liu Y L, Jin L W and Bai X. 2023. The development, application, and future of LLM similar to ChatGPT. Journal of Image and Graphics, 28(9): 2749-2762 (严昊, 刘禹良, 金连文, 白翔. 2023. 类ChatGPT大模型发展、应用和前景. 中国图象图形学报, 28(9): 2749-2762) [DOI: 10.11834/jig.230536]

Ye J B, Hu A W, Xu H Y, Ye Q H, Yan M, Dan Y H, Zhao C L, Xu G H, Li C L, Tian J F, Qi Q, Zhang J and Huang F. 2023. mPLUG-DocOwl: modularized multimodal large language model for document understanding [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2307.02499.pdf

Zhang Y Z, Zhang R Y, Gu J X, Zhou Y F, Lipka N, Yang D Y and Sun T. 2023. LLaVAR: enhanced visual instruction tuning for text-rich image understanding [EB/OL]. [2024-10-05]. https://arxiv.org/pdf/2306.17107.pdf

Zhong X, Tang J B and Yepes A J. 2019. PubLayNet: largest dataset ever for document layout analysis//Proceedings of 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney, Australia: IEEE: 1015-1022 [DOI: 10.1109/ICDAR.2019.00166]

作者简介

黎宇哲,男,硕士研究生,主要研究方向为计算机视觉。
E-mail: m202473925@hust.edu.cn
涂来,通信作者,男,副教授,主要研究方向为大数据挖掘与知识发现。E-mail: tulai@hust.edu.cn
伏凌,男,博士研究生,主要研究方向为计算机视觉。
E-mail: ling_fu@hust.edu.cn
朱泠皞,男,硕士研究生,主要研究方向为计算机视觉。
E-mail: linghaozhu@foxmail.com
罗琪頔,男,硕士研究生,主要研究方向为视觉与自然语言处理。E-mail: qdiluo@hust.edu.cn