中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：22721497

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	基于人类认知特征增强的视觉富文档理解方法研究

作　　者　王庆旋
学科专业　计算机应用技术
导　　师　王昊
完成日期　二〇二五年四月

姓　　名：王庆旋　　学号：22721497

论文题目：基于人类认知特征增强的视觉富文档理解方法研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名:

主　　席:

委　　员:

导　　师:

答辩日期: 2025 年 5 月 29 日

姓　　名：王庆旋　　学号：22721497

论文题目：基于人类认知特征增强的视觉富文档理解方法研究

上海大学学位论文原创性声明

本人郑重声明: 所呈交的学位论文是本人在导师指导下, 独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名: 王庆旋

日期: 2025年5月29日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即: 学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

(保密的论文在解密后应遵守此规定)

学位论文作者签名: 王庆旋　　导师签名: 王昊

日期: 2025年5月29日　　日期: 2025年5月29日

上海大学工学硕士学位论文

基于人类认知特征增强的视觉富文档理解方法研究

作　　者: 王庆旋
导　　师: 王昊
学科专业: 计算机应用技术

计算机工程与科学学院
上海大学
2025 年 4 月

A Dissertation Submitted to Shanghai University for the
Degree of Master in

Research on Human Cognition-Based Visually-rich Document Understanding Methods

Candidate: Wang Qingxuan
Supervisor: Wang Hao
Major: Technology of Computer Application

Schoool of Computer Engineering and Science
Shanghai University
April,2025

摘 要

随着信息化与数字化的迅猛发展，视觉富文档作为一种版式复杂、多模态融合且内容丰富的信息载体，广泛应用于日常生活以及金融、医疗、法律、物流等诸多领域的商业活动中。开发智能文档模型以实现视觉富文档的自动化处理具有重要的研究价值。传统文档智能模型通常依赖独立的 OCR 引擎处理文本信息、版式解析工具分析排版样式以及视觉编码器提取视觉特征，这种方式导致多模态信息难以有效融合。相比之下，人类读者通过多模态深度协同的认知机制，能够更高效地理解视觉富文档的内容。然而，目前针对这一多模态协同认知机制的研究较为匮乏，亟需深入探索。

先前的认知心理学研究通过深入探讨人类阅读理解过程发现，在理解视觉富文档时，阅读顺序与视觉兴趣区域的注意力机制对成功完成阅读任务起着至关重要的作用。阅读顺序是实现多模态信息融合的关键线索，它通过序列化处理文档，为多模态数据的分割、对齐与融合提供了重要的基础框架。此外，由于特定领域的视觉富文档通常具有独特的版式布局，人类在理解这些布局与文档结构时，往往依赖先验经验，通过注意力迁移快速定位关键区域，并借助注意力机制高效提取文档中的重要信息。因此，针对以上两点发现，本文首先利用眼动追踪技术构建了人类阅读视富文档的眼动数据集，在此基础上，围绕人类阅读文档时的阅读顺序与注意力机制在视觉富文档理解中的应用展开探讨，并开展以下研究工作：

(1) 基于眼动追踪与预排序模型的仿人类阅读顺序嵌入方法：当前的文档智能理解模型或多模态大模型通常直接输入 OCR 解析得到的文本和坐标信息，但边界框（Bounding Box）的乱序问题普遍存在。本研究进一步发现，无论是文档智能理解模型（如 LayoutLM 系列）还是多模态大模型，输入文本的顺序均会在一定程度上影响其在下游任务中的表现，尤其是在版面结构复杂的文档中。因此，本研究针对这一问题，重点探讨如何使文档智能模型学习人类阅读顺序，并验证是否通过融入人类阅读顺序能够提升视觉富文档的理解效果。具体而言，本研究利用眼动追踪仪采集人类阅读文档时的顺序信息，构建了一个包含人类眼动信息的基准数据集，并通过数据清洗与处理，将人类阅读顺序集成到文档数据中。此外，本研究受统计机器翻

译排序模型的启发，提出了一种模仿人类阅读顺序的预排序模型，基于文本、坐标、文本 + 坐标以及文本 + 坐标 + 图像四种模态生成仿人类阅读顺序，并将其应用于文档理解的下游任务。本研究不仅评估了生成阅读顺序与人类阅读顺序的相似性，还通过下游任务实验验证了其效果。实验结果表明，基于预排序模型的仿人类阅读顺序嵌入方法能够在一定程度上提升现有模型在下游任务中的表现，但也揭示了人类阅读顺序并非现有模型的最优解。

(2) 基于眼动追踪与热图提示的仿人类注意力机制学习方法: 人类读者通过视觉注意力等认知机制，能够高效筛选文档中的关键信息，优先处理高价值内容，并建立复杂的视觉语义关联。因此，本研究聚焦于人类注意力区域的先验经验对大模型在视觉富文档理解中的影响。我们利用眼动仪记录受试者的热力图等视觉行为数据，经过数据清洗和人工校验，构建了高质量的人类兴趣区域标注基准数据集。进一步地，我们将文档划分为不同数量的区域，通过颜色编码量化区域重要性，引导模型预测答案来源区域并生成注意力分数，同时结合提示学习完成下游任务。实验结果表明，视觉注意力与任务表现存在显著相关性，融入注意力热图提示后，模型在文档问答任务中的表现显著提升。研究还发现，大模型存在注意力对齐效应：尽管模型无需完全复制人类的视觉行为模式，但通过热图权重增强对高关注区域的聚焦，其任务表现能够显著提高。

综上所述，针对现有文档智能模型难以有效理解视觉富文档复杂结构的问题，本研究提出两种方法：一是以眼动追踪和预排序模型为基础的阅读顺序嵌入方法，通过模仿人类阅读顺序优化模型输入；二是基于眼动追踪和热图提示的注意力机制学习方法，通过强化模型对高关注区域的处理提升任务表现。研究结果发现，融入人类阅读顺序和注意力区域可显著提升模型在下游任务中的表现，尤其在复杂版面文档中。然而，人类阅读顺序并非模型最优解，大模型亦可以通过注意力对齐部分替代人类认知模式。以上发现一定程度上说明了在实现高效文档理解这一任务上，人工智能其实无需完全复制人类智能。本文的贡献在于不仅为文档智能模型的理解能力提升提供了新的解决方案，拓宽了眼动追踪技术在 AI 模型优化中的应用路径，也为认知心理学与计算机科学的跨学科研究，提供了新的研究范式和思路。

关键词：视觉富文档理解；眼动追踪；人类阅读顺序；注意力机制；大语言模型

ABSTRACT

With the rapid advancement of informatization and digitization, visually rich documents, characterized by complex layouts, multimodal integration, and rich content, have become ubiquitous in daily life and commercial activities across fields such as finance, healthcare, law, and logistics. Developing intelligent document models to enable automated processing of visually rich documents holds significant research value. Traditional document intelligence models typically rely on standalone OCR engines for text extraction, layout parsing tools for analyzing formatting styles, and visual encoders for capturing visual features, resulting in challenges in effectively integrating multimodal information. In contrast, human readers leverage deeply synergistic multimodal cognitive mechanisms to understand visually rich documents more efficiently. However, research on these multimodal cognitive mechanisms remains limited, necessitating further exploration.

Prior cognitive psychology studies have revealed that, during the comprehension of visually rich documents, reading order and the attention mechanisms directed toward regions of visual interest play critical roles in successfully completing reading tasks. Reading order serves as a key clue for multimodal information fusion, providing a foundational framework for segmenting, aligning, and integrating multimodal data through sequential document processing. Additionally, as visually rich documents in specific domains often feature unique layouts, humans rely on prior experience to understand these structures, rapidly locating key regions through attention shifts and efficiently extracting critical information via attention mechanisms. Building on these findings, this study first constructs a human eye-tracking dataset for visually rich document reading using eye-tracking technology. Subsequently, it explores the application of human reading order and attention mechanisms in understanding visually rich documents, conducting the following research:

(1) Human-like Reading Order Embedding Method Based on Eye Tracking and Pre-sorting Models: Current document intelligence models or multimodal large models typically input text and coordinate information parsed by OCR, but the issue of disordered

bounding boxes is prevalent. This study finds that the order of input text affects the performance of both document intelligence models (e.g., LayoutLM series) and multimodal large models in downstream tasks, particularly for documents with complex layouts. To address this, the study investigates how to enable document intelligence models to learn human reading order and verifies whether incorporating human reading order enhances the comprehension of visually rich documents. Specifically, human reading order data is collected using eye-tracking devices, and a benchmark dataset incorporating this eye-tracking information is constructed through data cleaning and processing. Inspired by statistical machine translation reoredering models, this study proposes a pre-reoredering model that mimics human reading order, generating human-like reading sequences based on four modalities: text, coordinates, text+coordinates, and text+coordinates+images. These sequences are applied to downstream document understanding tasks. The study evaluates the similarity between generated and human reading orders and validates their effectiveness through downstream task experiments. Results show that the human-like reading order embedding method based on pre-sorting models improves existing models' performance in downstream tasks to some extent but also indicates that human reading order is not the optimal solution for current models.

(2) Human-like Attention Mechanism Learning Method Based on Eye Tracking and Heatmap Hints: Human readers, through cognitive mechanisms such as visual attention, efficiently filter key information, prioritize high-value content, and establish complex visual-semantic associations. This study focuses on the impact of prior human attention region experience on large models' understanding of visually rich documents. Using eye-tracking devices, we record subjects' heatmap and other visual behavior data, constructing a high-quality annotated benchmark dataset of human regions of interest through data cleaning and manual validation. Furthermore, documents are segmented into varying numbers of regions, with region importance quantified via color encoding to guide models in predicting answer source regions and generating attention scores, combined with prompt learning for downstream tasks. Experimental results demonstrate a strong correlation between visual attention and task performance, with significant improvements in document question-answering tasks after incorporating heatmap prompts. The study also identifies an attention

alignment effect in large models: while models do not need to fully replicate human visual behavior, enhancing focus on high-attention regions through heatmap weights significantly boosts task performance.

In summary, to address the challenge of existing document intelligence models struggling to effectively understand the complex structures of visually rich documents, this study proposes two methods: (1) a reading order embedding method based on eye-tracking and pre-sorting models, which optimizes model input by mimicking human reading order; and (2) an attention mechanism learning method based on eye-tracking and heatmap prompts, which enhances task performance by strengthening model focus on high-attention regions. The results show that incorporating human reading order and attention regions significantly improves model performance in downstream tasks, particularly for documents with complex layouts. However, human reading order is not the optimal solution, and large models can partially replicate human cognitive patterns through attention alignment. These findings suggest that, for efficient document understanding, artificial intelligence does not need to fully mimic human intelligence. This study' s contributions include providing new solutions for enhancing document intelligence models, broadening the application of eye-tracking technology in AI model optimization, and offering a novel interdisciplinary research paradigm for cognitive psychology and computer science.

Keywords: Visually rich document understanding; eye-tracking; Human reading order; Attention mechanism; Large language models

目　录

第一章　绪论

1.1　研究背景和意义

视觉富文档[1]是包含丰富视觉元素的文档，这些元素不只是文本，以及图像、表格、排版等多种形式的信息，它们可相互补充，在数字化潮流的推动下，信息载体从传统的纯文本演变成了包含丰富视觉元素的富文档，像 PDF、网页、扫描图像等。如图 1.1所示，这些文档有文字信息，还融合了表格、图像、布局、字体、颜色等多种视觉元素，承载着更丰富复杂的语义信息，不过传统的 OCR 技术主要针对纯文本图像，很难有效处理富文档里复杂布局、表格、图像等非文本信息，成了制约信息提取效率和智能化应用发展的瓶颈。

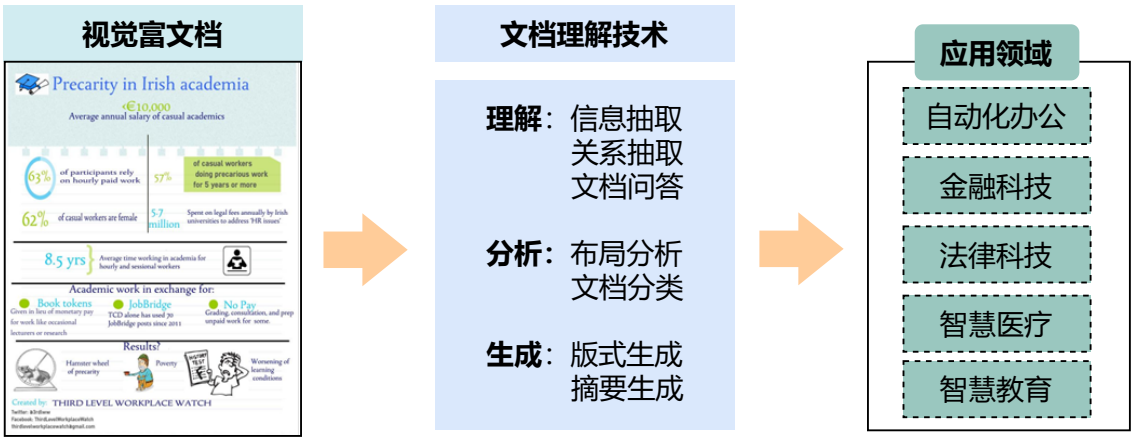


图 1.1　文档智能理解技术的应用场景

近些年，深度学习、自然语言处理、计算机视觉等[2]人工智能技术快速发展，给视觉富文档理解提供了新的解决办法。视觉富文档理解研究利用人工智能技术，自动解析文档结构并提取文本、表格、图像等多模态信息，有关键的理论意义和实际应用价值，这些任务包含文档分类、信息抽取、表格识别、版面分析等，是要把非结构化的文档内容转化为结构化的数据，来支持信息检索、知识管理、智能办公等应用。视觉富文档理解的核心挑战在于处理复杂的版式结构、顺序结构、多样的视觉特征以及低质量的扫描图像，以达成准确且高效的文档语义理解，该研究能有效提升信息提取效率，推动文档分类、信息检索、知识图谱构建等智能化应用发展，还可

以拓展人工智能在金融、医疗、法律和教育等领域的应用场景，比如合同分析、病历分析、法律文书处理和自动化作业批改等，有广阔的应用前景。然而跟纯文本文档相比，视觉富文档因其复杂的版式结构和多样的视觉特征更具挑战性，文档扫描图像质量不佳和文档形式不统一等问题，也加大了视觉富文档语义理解的难度。

未来，视觉富文档理解研究将朝着以下几个方向发展：Goodfellow 等人[2]在多模态信息融合中提出如何有效地融合文本、图像、布局等多模态信息，提升文档理解的准确性；复杂文档结构解析[3]，针对表格、公式、流程图等复杂文档结构，研究更加鲁棒和高效的解析算法；领域自适应[4]，研究如何将视觉富文档理解技术应用于特定领域，例如医疗、金融等，提升领域文档的理解效果；人类认知特征的集成，通过加入人类认知让机器学习人类的理解文档的方式，提升机器理解的效果。

视觉富文档理解研究是一个充满挑战和机遇的领域，其发展将推动人工智能技术在文档处理领域的应用，为信息提取、知识管理、智能办公等领域带来变革，具有重要的社会和经济价值。

1.2 研究问题

本文研究的对象是视觉富文档和人类认知特征，由于研究对象的结构布局的复杂以及多模态无法对齐等特点，现有的文档智能模型往往只关注模型的本身，而忽略了在模型中加入人类的认知特征。因此，本文主要关注以下两个问题：

(1) 如何模仿人类阅读顺序并能够改善模型对复杂版式文档的理解能力？

研究发现，由于文档的多样性和复杂性，识别和理解视觉丰富的文档是一项费时费力的工作。单纯的文本信息不足以从不同的文档类型中提取关键信息，这就需要一种多模态的方法，通过联合建模考虑多种模式的一致性和相关性，包括文本、视觉和布局。用于理解 VRD 的现代文档人工智能模型的例子包括 LayoutLM[5-7]系列模型和 StrucText[8-10]系列模型。虽然这些模型可以获得细粒度的多模态文档表征，并在下游的视觉富文档理解任务中取得显著效果，但它们缺乏从给定文档中产生适合 Transformer[11]架构的序列化输入顺序的能力。

因此，它们通常利用简单的规则，有时直接使用 OCR 工具在前一步生成的输入顺序。然而，这些输入顺序与人类习惯的阅读顺序有很大不同，人类的阅读顺序通常遵循一定的逻辑和视觉线索，例如从左到右、从上到下，以及根据文档的排版结

构（如标题、段落、表格等）进行信息获取。具体而言，人类在阅读时会自然地根据文档的布局和视觉线索调整注意力分布，例如优先阅读标题、图表说明或加粗文本，而现有模型往往缺乏这种动态调整能力。此外，人类能够通过上下文快速理解视觉元素（如图表、表格）与文本之间的关系，而现有模型在处理多模态信息时，往往存在信息割裂的问题。这些局限性限制了模型在复杂文档理解任务中的表现。通过将阅读顺序特征融入文档智能模型，可以帮助模型更好地理解文档的层次结构和语义关系，尤其是在处理包含复杂布局、多栏排版或混合文本与图像的文档时。阅读顺序的引入能够引导模型按照更符合人类认知的方式处理信息，从而提升对文档内容的整体理解效果。不同的输入顺序会大大影响文档人工智能模型在下游文档理解任务上的表现，这一点常常被忽视。所以针对上述问题，如何构建出正确的阅读顺序，提升视觉富文档的理解能力，是需要解决的问题。

(2) 如何使大语言模型通过提示感知学习人类注意力机制？

文档智能技术的快速发展显著改变了我们获取、处理和解析信息的方式[1]。富视觉文档（VRD）因其包含表格、图形、图表等多种视觉元素，能够承载大量信息，在各个领域得到了广泛应用。然而，理解这类文档不仅需要处理文本数据，还需要整合视觉元素、布局结构，并结合与人类感知和认知过程相匹配的机制。因此，Chen[12]等人提出利用视觉-排版跨模态提示来提升多模态视觉富文档模型的理解能力。与此同时，人类在阅读富视觉文档时，依赖于阅读顺序、视觉注意力等认知机制，能够高效地提取信息。这些机制使读者能够聚焦于文档的特定区域，优先处理关键信息，并理解复杂的视觉上下文关系。

相比之下，尽管大语言模型（LLMs）和视觉大语言模型（VLLMs）在基于人工智能的文档理解领域取得了显著进展，但这些模型在模拟人类认知深度方面仍存在明显不足，尤其是在如何利用视觉注意力机制来增强内容理解方面[13]。人类的认知机制是长期进化形成的，使个体能够在阅读过程中有效区分信息优先级，并将文本、视觉和布局数据无缝整合。然而，现有的 LLMs 和 VLLMs 虽然在文本理解和生成任务中表现出色，但在处理富视觉文档时，仍难以完全复现人类的认知能力。因此，进一步探索如何将人类的视觉注意力融入文档智能模型中也是一个急需要解决的问题。

1.3 研究内容

基于以上研究问题和人类认知心理学发现，本文要解决将人类阅读顺序注入到模型中和让大语言模型感知通过人类注意力以获取视觉富文档重点区域来提升对视觉富文档理解的能力，如图 1.2所示。

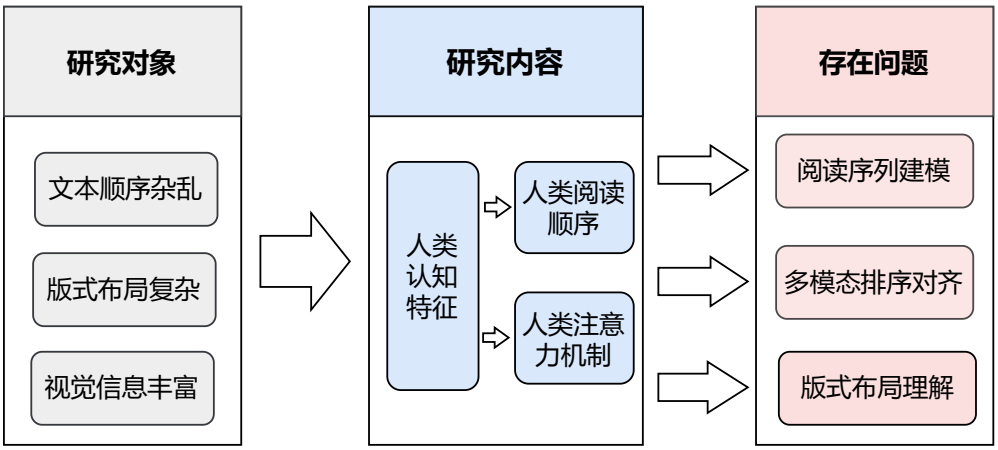


图 1.2　本文的研究内容

（1）为了解决文档版式复杂阅读顺序紊乱导致的无法对阅读序列建模问题，我们提出了加入人类阅读顺序和仿人类阅读顺序来学习人类的阅读顺序以提升机器对视觉富文档的理解能力。首先我们通过眼动追踪记录仪对已有的视觉富文档数据集进行标注，再将标注的人类阅读顺序提取出来，并对缺失点、重复点和边界点进行处理获取每个文本框的顺序。然后我们根据文本、坐标、文本 + 坐标和文本 + 坐标 + 图像四种不同模态的特征对文本框进行排序，四种预排序模型通过和人类阅读顺序做对比损失来学习适合机器的阅读顺序，并评估了预排序模型的顺序和人类阅读顺序的相似性。除此之外，我们还提出一种基于“Z”字型顺序的模式，即只通过坐标对每个文本框进行规则性的排序。最后我们将排序后的结果输入下游任务的模型中进行评估，实验结果表明了我们提出的预排序方法有效性，同时也证明了人类阅读顺序可以提升模型对视觉富文档理解的能力。

（2）为了解决视觉富文档理解中大模型对重点区域关注度不足导致的版式布局理解困难的问题，我们提出了一种新的方法，通过模拟人类注意力机制，使大语言模型能够更好地关注视觉富文档中的关键区域。具体而言，我们首先利用眼动追踪技术收集人类在阅读视觉富文档时的注意力数据，包括注视点、注视时间和扫视路径等信息。然后，我们设计了一种注意力对齐机制，将人类的注意力数据与模型的注

意力机制进行对齐，使模型能够学习到人类在阅读过程中的注意力分配模式。此外，我们还提出了一种注意力相似性评估的方法，通过将文档切分成不同的区域，根据颜色对每个区域进行打分，然后再让大模型预测答案最有可能哪个区域以获取大模型的注意力分数。实验结果表明，大模型获得的注意力分数在下游任务中的分数也越高，证明了我们的方法能够显著提高大语言模型在视觉富文档理解任务中的性能，特别是在识别和理解关键区域方面表现出色。

1.4 创新点

本文主要针对视觉富文档理解任务，从人类认知心理学的角度出发，对人类阅读顺序和人类注意力机制两大方面进行研究，在图 1.3展示了本文的创新点。

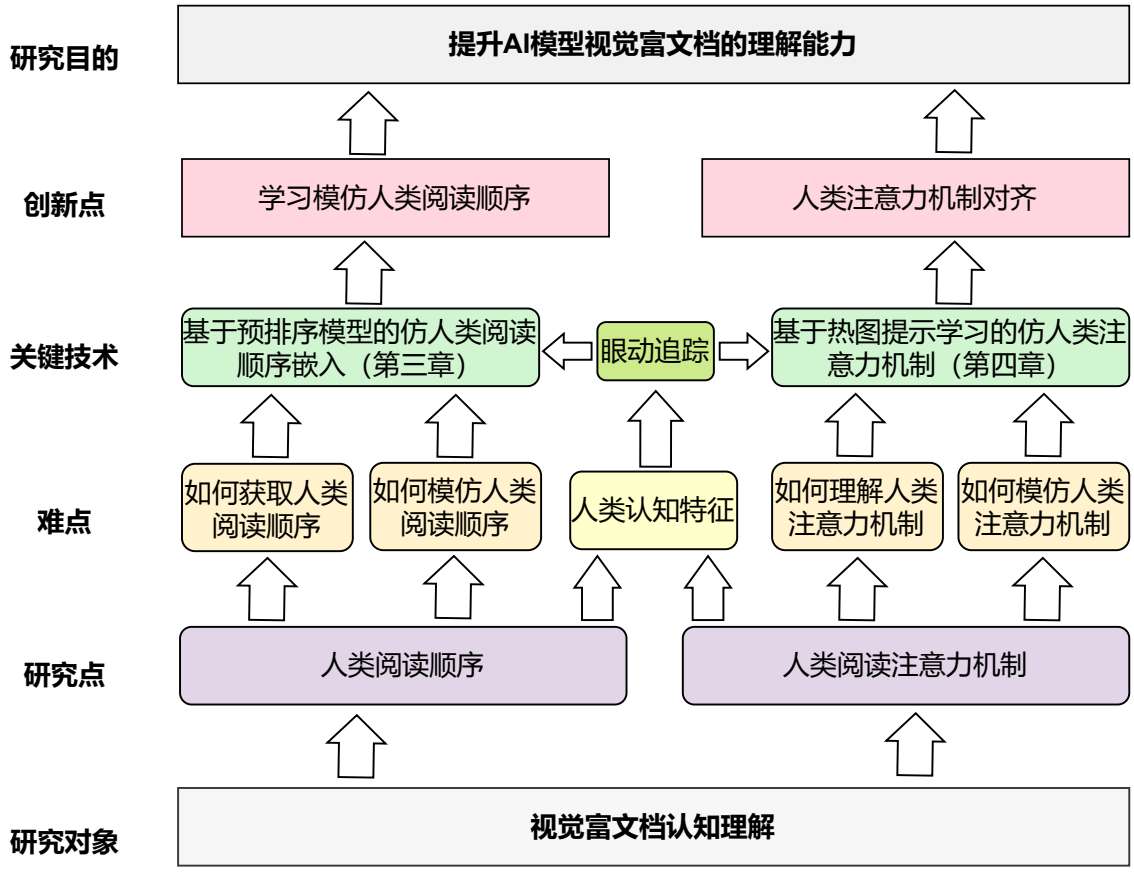


图 1.3　本文的创新点

（1）提出多模态预排序模型，并在预排序模型中融合了人类阅读顺序，实现阅读顺序和多模态特征的对齐：第一个创新点是通过注入预排序实现人类阅读顺序的融

合。眼动追踪技术可以获取人类阅读时的眼动特征，通过这种技术获得了人类阅读顺序；预排序是一种对文档内容进行预先处理和排序的技术，其目的是让 AI 模型能够按照更接近人类习惯的顺序来理解和处理视觉富文档，解决了如何模仿人类阅读顺序这一难点。本创新点的关键技术是基于预排序模型的仿人类阅读顺序生成，通过这种方式，AI 模型可以更好地模拟人类在阅读文档时的顺序和逻辑，从而提升其对文档内容的理解能力。除此之外，我们将模仿出的人类阅读顺序与人类阅读顺序进行相似性评估，并在下游任务中证明了本创新点可以解决本研究内容的难点。

（2）提出了通过热图使用大语言模型感知人类注意力机制，实现大模型在视觉富文档理解时对重点区域的关注：在研究过程中，如何理解人类注意力机制以及如何模仿人类注意力机制是需要深入探讨的难点问题。我们的创新点是提出基于眼动追踪的人类注意力建模方法，通过采集注视点、扫视路径等细粒度数据，构建视觉富文档的注意力分布热图，并设计跨模态注意力对齐机制，使大语言模型能够学习人类阅读时的动态注意力分配模式。进一步提出区域化注意力相似性评估框架，将文档切分为不同数量的区域，通过颜色编码量化区域重要性，引导模型预测答案来源区域并生成注意力分数。基于热图提示学习的仿人类注意力机制是实现这一创新点的关键技术，它使 AI 模型能够感知并学习人类在阅读文档时的注意力分配模式，从而更准确地理解文档的重点和细节。实验表明，注意力分数与下游任务性能呈强正相关，验证了模型对人类注意力机制的精准模仿。技术突破点在于解决“如何从眼动数据中提取鲁棒的注意力特征”以及“如何通过区域化评分实现注意力可解释性对齐”两大挑战，显著提升模型对文档关键区域的定位与理解能力。

1.5 本文的组织架构

本文针对人类认知特征对视觉富文档理解展开研究，本文首先介绍了视觉富文档理解的研究背景和意义，已有的视觉富文档理解模型、视觉富文档理解中的任务以及人类认知特征。本文主要研究了人类阅读顺序和模仿人类阅读顺序对视觉富文档的理解以及人类注意力区域在大模型中对 DQA 任务的影响，最后总结了本文的研究工作并进行展望。

第一章绪论作为文章的开篇，主要介绍视觉富文档理解的研究背景和重要性。将概述视觉富文档理解所面临的挑战和现有研究的局限性，进而引出本研究旨在解决

的问题和研究的创新点。

第二章将深入探讨视觉富文档理解的相关工作，首先介绍了视觉富文档理解任务的定义；紧接着介绍了传统模型和大语言模型在视觉富文档理解中的发展脉络，比较了现有方法的不足；最后，介绍了眼动追踪技术和人类认知特征相关内容。

在第三章中，研究将聚焦于基于人类阅读顺序的视觉富文档理解方法。本章将比较机器顺序和人类顺序，探索如何模拟人类的阅读习惯来提高文档理解的准确性。通过研究，旨在提出一种新的方法，使机器能够更好地理解和处理视觉富文档。

第四章将基于人类注意力机制，研究视觉富文档理解方法。本章将探讨如何通过划分注意力区域和评测注意力相似性来模拟人类的注意力分配。研究的目的是开发出能够更自然地处理文档的模型，这些模型能够识别和专注于文档中的关键信息。

第五章将对全文进行总结，并对未来的研究方向进行展望。本章将回顾前述章节的主要发现和贡献，并讨论研究的局限性和未来改进的可能方向。通过总结，本文旨在为视觉富文档理解领域提供新的视角和研究思路。

第二章　视觉富文档理解方法和人类认知特征研究

2.1　相关任务定义

视觉富文档理解（Visually-rich Document Understanding，VrDU）是一项涉及多模态数据处理和智能分析的技术[14]，旨在从包含丰富视觉元素的文档中提取、解析和理解文档信息。这类文档通常包括 PDF、Word、HTML/XML 等格式，内容涵盖文本、图像、表格、图表等多种元素。视觉富文档理解的核心任务是从这些复杂的多模态数据中提取有用的信息，并将其转化为结构化的知识，以支持各种下游应用。

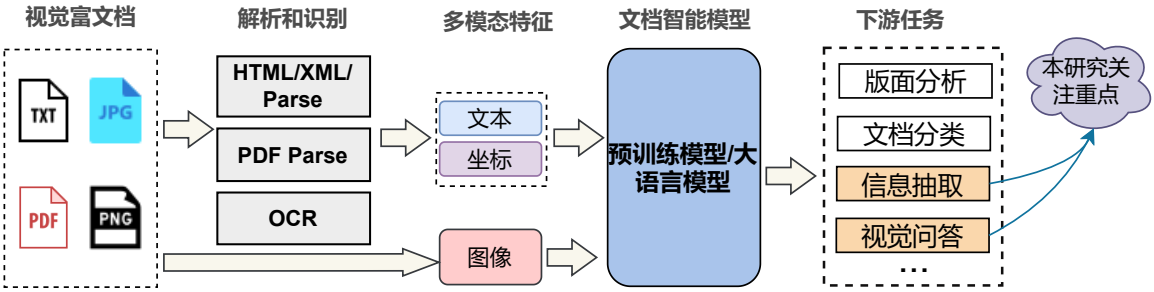


图 2.1　现有的视觉富文档理解技术框架

视觉富文档理解的第一步是对文档进行解析与识别，以提取其中的多模态特征。这一过程通常包括以下几个关键步骤：

1）**PDF 解析**：从 PDF 文档中提取文本、图像、表格和布局信息。PDF 文档通常包含复杂的版式结构，解析时需要识别文档中的标题、段落、列表、表格等元素。

2）**OCR（光学字符识别）**：对于扫描图像或包含图像内容的文档，OCR 技术用于将图像中的文字转换为可编辑的文本。OCR 的准确性直接影响到后续文本处理的效果。

3）**多模态特征提取**：从文档中提取文本、坐标、图像等多模态特征。文本特征包括文字内容及其语义信息，坐标特征用于描述文本和图像在文档中的位置关系，图像特征则包括图表、图形等视觉元素。

在解析与识别的基础上，通过文档智能模型对提取的视觉富文档的多模态特征进行进一步处理和分析。这些模型通常基于预训练的语言模型或大语言模型，能够对文档内容进行深度理解和表示。预训练模型/大语言模型：利用预训练的语言模型

或视觉大语言模型对文档内容进行编码和表示。这些模型能够捕捉文本、图像和布局之间的复杂关系。

文档智能模型的应用场景广泛，涵盖了多个下游任务，主要包括：

- 版面分析[15]：识别文档的版面结构，如标题、段落、表格、图像等。版面分析是理解文档内容的基础，能够帮助模型更好地处理复杂的文档布局。
- 文档分类：根据文档内容将其归类到预定义的类别中。例如，将文档分类为合同、发票、简历等类型，以便后续处理。
- 信息抽取[16]：从文档中提取特定的信息，如实体、关系、事件等。信息抽取是视觉富文档理解的核心任务之一，能够将非结构化的文档内容转化为结构化的数据。
- 文档问答：根据文档内容回答用户提出的问题，如图 2.2左图所示。视觉问答任务要求模型能够理解文档中的文本和图像信息，并结合上下文生成准确的答案。

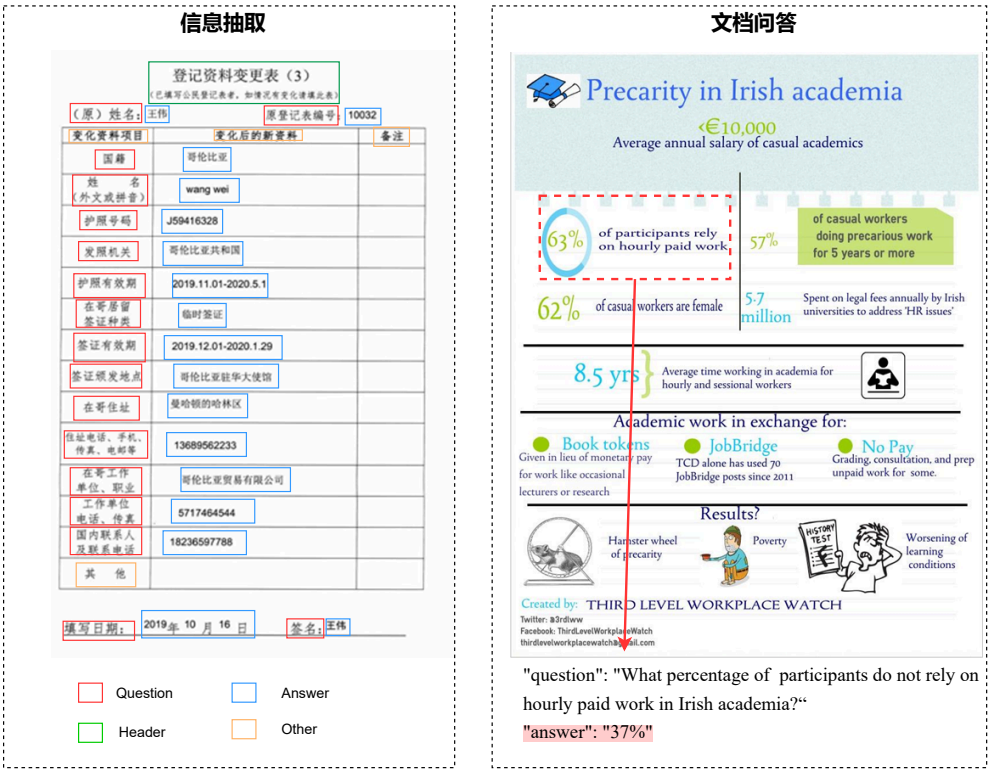


图 2.2　信息抽取任务和文档问答任务

视觉富文档理解的任务定义涵盖了从文档解析、多模态特征提取到智能模型应用的完整流程。其核心目标是通过整合文本、图像和布局等多模态信息，实现对文

档内容的全面理解和高效处理。未来，随着人工智能技术的不断发展，视觉富文档理解将在更多领域发挥重要作用，推动文档处理技术的智能化和自动化。

2.2 视觉富文档理解方法

2.2.1 文字识别方法

光学字符识别（Optical Character Recognition, OCR）是一种通过扫描和处理图像，将图像或书面文档中的文字提取为可编辑、可搜索的数字文本的技术。其主要作用是将印刷体或手写体的图像等转换成计算机可识别的字符，最终实现将图像转化为数据。OCR 技术在多个领域有广泛应用，例如文档数字化：将纸质文档转换为易于存储和检索的电子文本；金融行业：用于证件识别、身份验证等场景，提高效率和准确性；教育领域：通过扫描试卷、兑现赠款等方式提高教育质量；物流领域：用于确定快件号码、记录货物信息；医疗领域：电子病历的使用及医疗报告的分析。

（1）早期阶段：基于模板匹配的 OCR

早期的光学字符识别技术一般是基于模板匹配的技术，其核心思想是将待识别的字符图像与一个或多个预定义的模板进行对比，找到与之最相似的模板，然后输出对应的字符[17]。这种方法首先构建一个包含所有可能字符的模板库，然后将输入的文档图像分割成单个字符图像，并与模板库中的模板进行比对，选择相似度最高的模板作为识别结果[18-19]。

利用模板匹配的 OCR 识别方案虽然结构简单、直观，识别准确度高且速度快，用来识别任何标准化印刷字符都非常合适，但它非常局限于字体和字号，不可能应用到各种不同的情况中。而且它对噪声很敏感，不具有处理手写体的能力，尤其是它的模板库只能有有限的几张图片存在，在现在信息爆炸的时代更显得无能为力。也正是这样的缺陷让传统的基于模板匹配的 OCR 方案难以适应如今变化日新月异的节奏。

虽然基于模板匹配的 OCR 处理取得了不错的效果，但是日益增长的字符识别需求也揭示出了它的弊端。这一阶段的工作为之后 OCR 的发展做了铺垫，并且启发人们应用更优秀、更先进的统计及深度学习模型到 OCR 中来，使 OCR 系统能够适应更多的复杂字体、多种背景和多种语言等其他任务。

（2）统计方法阶段：基于特征提取的 OCR

随着统计学和机器学习领域的兴起，OCR 对于特征的依赖越来越强[20]。将一个字符的图像经过特征提取之后得到一些形状、边缘轮廓及笔画等可视特征，并利用这些特征通过一定形式的支持向量机或隐马尔科夫模型进行分类的过程均属于该种方法。此时期应用 OCR 采用的人工设计特征的方法居多，在此基础上运用统计数据、各种机器学习模型实现字符识别。这一时期的 OCR 要求人们自行获取有效的特征值并形成特征空间代表待识别人文数据中的字符从而达到识别的目的。对每一个需要被判断的人物画像来说，找到合适的、有效反映该人物画像本质的特征至关重要，这也是这个步骤的核心所在。而特征提取在整个 OCR 工作过程中发挥着举足轻重的作用，常用的提取特征技术包括以下几种经典方法：一是基于结构形态的特征提取，如边界特征法和傅里叶特征算子，通过检测字符的边界或频域特征来提取信息[21-22]。二是基于几何分布的特征提取，如二维直方图投影法和区域网格统计法，通过分析字符的几何分布和局部特征来提取信息[23-24]。三是基于统计量的特征提取，如灰度直方图和纹理特征[25]，通过统计图像的灰度分布或纹理结构来提取信息。

图像特征提取完成后，OCR 技术通过机器学习算法对特征进行分类和识别。常用的特征分类器有支持向量机（SVM）、隐马尔可夫模型（HMM）和条件随机场（CRF）[26-28]。这些分类器能够根据提取的特征对字符进行分类，同时考虑字符之间的上下文关系，进而实现将识别的准确性提高。通过学习大量带标签的训练数据，构建字符的分类器模型。这些方法的优势在于可以在一定程度上适应字符的变化，并能够识别手写体和复杂字体。

（3）深度学习阶段：基于神经网络的 OCR

随着深度学习技术的来临，基于神经网络的 OCR 技术逐渐成为主流。这一阶段的技术通过卷积神经网络（CNN）[29]、循环神经网络（RNN）[30]、注意力机制等模型，实现了从图像中自动提取特征并进行文本识别，显著提升了 OCR 的准确性和鲁棒性。

- **卷积神经网络（CNN）**：用于提取图像中特征的 CNN 是深度学习 OCR 中的核心组件。通过卷积层、池化层和全连接层的组合，CNN 能够自动学习图像中的局部和全局特征。在 OCR 任务中，CNN 通常用于两个阶段：文字检测（识别图像中的文字区域）和文字识别（从检测到的区域中提取文字内容）。例如，在古籍中，CNN 能够有效捕捉文字的笔画、字形结构和纹理等特征，从而提高识别精度[31]。

- **循环神经网络（RNN）及其变体**：与 CNN 有所不同，RNN 及其变体（如 LSTM 和 GRU）主要用于处理序列数据[32-33]，尤其是文本行中的字符序列。在 OCR 中，RNN 能够捕捉字符之间的上下文关系，从而提高识别的准确性。

2.2.2 基于多模态预训练方法

对于 NLP 任务而言，使用大规模无标注语料进行自监督预训练取得了不错的成果。BERT[34] 模型通过 MLM (Masked Language Model) 的方式实现；GPT[35] 则是利用 NTP（Next Token Prediction）的方式实现类似的功能。基于这些方法，人们也试图把它们用在对文档图像的预训练上，尤其是希望能够在大量的、未标注的文档图像上做预训练。如图 2.3 所示，基于 Transformer 架构的 VrDU 预训练技术分为两部分：首先是数据准备，需要选择预训练的数据源以及使用 OCR 技术提取视觉富文档的布局和文本信息等预处理操作；然后就是模型预训练，VrDU 预训练模型在此过程中会经历三个步骤：单模态表示，使用编码器获得文本、布局和图像多模态特征；多模态融合，将这三种模态相加或者拼接后用 Transformer 进行融合；最后是预训练任务，针对每种模态设定相应的预训练任务，发掘不同模态之间的关系和互补性，用于指导预训练。经过预训练后，待微调时它已经具备了可以有效捕捉到含语义丰富的文档表示的能力。

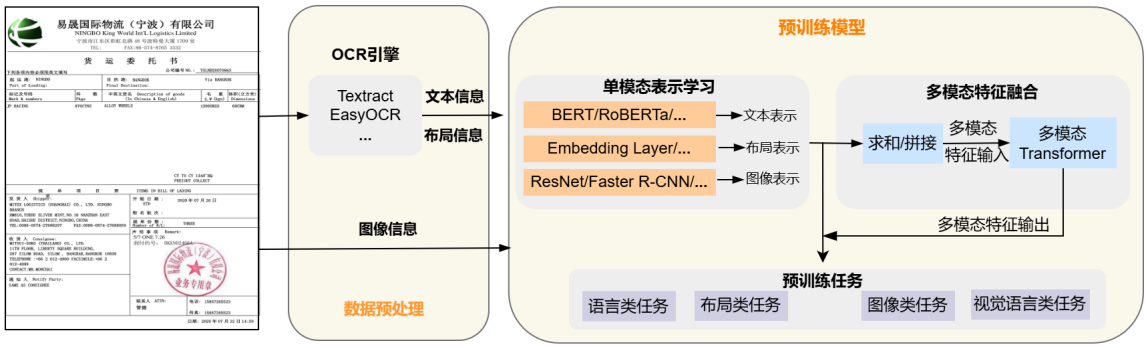


图 2.3　基于 Transformer 的预训练技术架构

BERT[34]：BERT 模型通过使用 WordPiece 构建了一个包含 30,000 个 token 的词汇表，并学习相应的文本语义表示。在句子中，第一个 token 为 [SEP]，并通过 [PAD] 进行填充，以确保每个句子序列的 token 长度保持一致。BERT 模型的文本语义表示通常是通过三种嵌入（token 表示、一维位置表示和片段编号表示）的加和来获得。

在大多数场景下，文档中字符的位置关系蕴含着丰富的语义信息。例如，表单

通常以键值对的形式呈现。一般而言，键值对的排列方式通常为左右或上下，并且存在特定的类型关系。对于富文本文档而言，除了文字本身的位置关系外，文字格式所呈现的视觉信息同样能够为下游任务提供支持。

LayoutLM[5]：为了解决这些问题，微软提出了文档预训练模型 LayoutLM。作为第一个大规模文档理解训练模型，LayoutLM 结合了文本中的语义和布局信息，并将这两种特征组合起来输入到 Transformer 编码器中。该模型基于 BERT 框架，在单个框架中同时集成文本和 2D 位置坐标，并结合了两个新的嵌入功能：2D 位置嵌入和图像嵌入。通过 Transformer 的自注意力机制，该模型可以输入文档中所有文本、坐标和图像信息，从而更高效地将图式信息与语义数据对齐，全面理解文档内容。对于各个层次文本的视觉属性，例如形状属性，例如大小、细度、颜色和倾斜度，这些特征为系列中的标记任务提供了重要信息。此外，LayoutLM 设计了两个预训练任务：视觉-语言建模和文档分类。当模型应用到最终的任务中时，Faster R-CNN 所获取的图像特征将会融入到最终的单词表达中。LayoutLM 作为视觉文档理解领域的先驱性研究，开辟了一条新的研究道路。将图像与文本特征相结合有助于增强文档的语言表达能力。在实际微调过程中，图像信息的引入并未显著提升下游任务的表现。

StructuralLM[36]：该模型在 LayoutLM 模型的基础上进行了 3 个方面的优化: 加入对单词的一维全局位置信息的建模, 丰富了空间布局信息的多样性; 同一个文本区域内的单词共享同样的二维全局位置信息，有利于单词更好地获取同一区域内的上下文信息。StructurallM 模型预训练任务采用了布局类和视觉语言类任务，其中布局类的单元位置分类 (Cell Position Classification,CPc) 采用粗粒度的方式来挖掘文本单元的布局信息。它们先将文档图像分割为多个大小一样的区域, 并根据二维空间位置计算文本单元属于哪一个区域；然后随机遮掩掉 15% 的文本单元的二维空间位置 (如左上点、右下点和中心点坐标等): 最后基于多模态特征和一个分类层来预测这些遮掩掉的文本单元属于文档图像中的哪一个区域。利用单元位置预测任务来替代多标签文档分类任务，该预训练任务通过预测特定文本区域在文档图像中所处的位置来对其相对位置信息进行建模。

尽管以上模型在文档理解任务中取得了一定成果，但存在以下局限性：其一，以上模型在 BERT 基础上加入了 x 轴和 y 轴的坐标信息，但是未充分考虑序列位置与二维空间的差异，导致模型难以有效融合并利用空间信息，在处理特定文档时表现欠佳；其二，预训练策略与 BERT 类似，而 BERT 未明确考虑文本块间的位置关系，

这使得模型可能无法充分利用文本块间的位置信息，从而限制了其在实际任务中的性能发挥；其三，以上模型忽略了多模态之间的对齐，只是将单模态进行简单的相加处理，未能真正实现多模态的间的融合。

LayoutLMv2[6]：LayoutLMv2 是微软研究院提出的第二代多模态预训练模型，旨在处理视觉信息丰富的文档。该模型在 LayoutLM 的基础上进行了优化，主要通过将图像信息融入预训练过程，以增强模态特征的多样性。此外，LayoutLMv2 设计了一种基于空间感知的自注意力机制，能够有效挖掘二维空间中的布局信息。模型采用文本-图像对齐任务替代了多标签文档分类任务，目的是深入探讨文本与图像之间的关系。最重要的是，该模型还考虑了区分图像和文本的模态类型编码，并为每种模态引入了一维全局位置编码。LayoutLMv2 模型首先对文本语义表示、图像语义表示及其对应的布局表示进行求和，然后将这两种表示连接起来，作为 Transformer 编码器的输入。与 LayoutLM 相比，LayoutLMv2 在性能上表现更为优越，尤其是在对视觉信息敏感的任务中，如文本-图像匹配和结构化数据提取等方面。

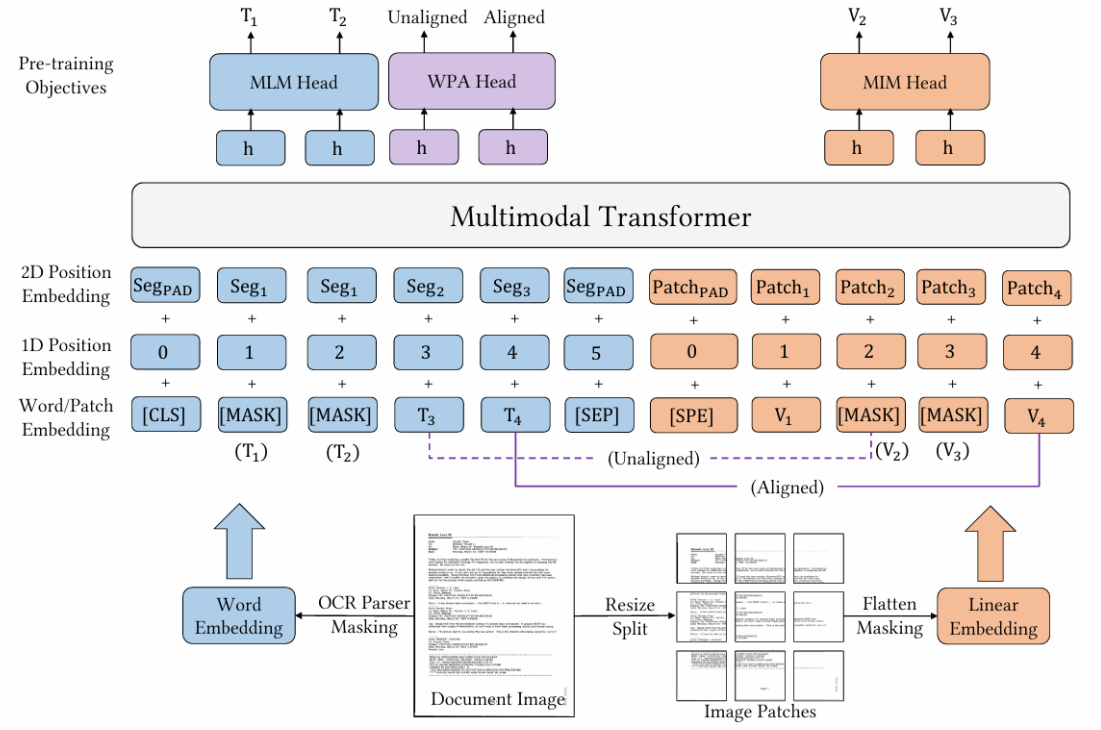


图 2.4　LayoutLMv3 的整体框架[7]

LayoutLMv3[7]：LayoutLMv3 是 LayoutLM 系列的第三代模型，融合了文本、布局及图像信息，标志着文档智能领域首个不依赖于预训练卷积神经网络（CNN）来提取视觉特征的多模态模型。与 LayoutLM 和 LayoutLMv2 采用的单词级布局位置嵌

入不同，LayoutLMv3 引入了段落级布局位置嵌入。这一设计的原因在于，同一段落内的词语通常具有相似的二维空间位置，并且往往承载相似的语义信息。在图像嵌入方面，该模型首先将图像划分为一系列均匀的块，然后将这些图像块线性投影到潜在向量维度，并展平为一系列向量。此外，LayoutLMv3 通过词块对齐任务进行预训练，旨在通过预测文本词对应的图像块是否被遮蔽来学习跨模态对齐关系。这种简洁而统一的架构及训练目标使 LayoutLMv3 成为一个通用的预训练模型，适用于以文本为中心和以图像为中心的智能文档处理任务，例如文档布局分析和视觉信息提取。

然而，以上大多数方法要么仅关注文本、布局和图像特征之间的融合，要么将边界框的位置关系作为预训练任务来建立边界框之间的联系，忽视了前后边界框的顺序关系，从而未能充分考虑整个边界框的顺序，导致模型在建模文档整体顺序方面的能力不足。

LayoutReader：Wang 等人[37]提出一种用于文档阅读顺序预测的模型 LayoutReader，它利用序列到序列（seq2seq）模型来编码文本和布局信息，并生成阅读顺序的索引序列。LayoutReader 模型结构包括编码器（Encoder）和解码器（Decoder）。在编码阶段，LayoutReader 将源序列和目标序列打包成一个连续的输入序列，并设计了自注意力掩码来控制 token 之间的可见性。具体来说，LayoutReader 允许源序列中的标记相互关注，同时阻止目标序列中的标记关注右侧上下文。在解码阶段，由于源序列和目标序列是重新排序的序列，预测候选可以被限制在源序列内。因此，模型被要求预测源序列中的索引。该模型通过结合文本内容和布局信息，提高了阅读顺序检测的准确性，这对于复杂文档的阅读顺序检测至关重要。

XYLayoutLM：Gu 等人[38]提出的是一种多模态文档理解模型，旨在通过获得合理的文本阅读顺序和提出空洞条件位置编码，该模型的输入包括图像视觉特征、文本特征和文本位置特征。视觉特征通过 ResNeXt-101[39]的特征图池化到 7x7 的特征图拉平得到，同时两个位置编码生成器把输入文本框编码成 pos embeddings 和 box embeddings。这些 embeddings 拼接起来后输入具有自注意力机制的 transformer 层，输出的视觉/文本 token 表征被用于文档理解任务。如图 2.5所示，与基准模型 LayoutXLM[40]不同，XYLayoutLM 通过 Augmented XYCut 的方法来获得正确的阅读顺序，同时使用了空洞条件位置编码模块去处理不同长度的序列。该模型解决了 LayoutLMv2 的两个局限性: 依赖 OCR 产生的 tokens 和 bbox，没有去探索阅读顺序带来

的影响; 通常使用固定长度的相对或者绝对位置编码，不能处理比固定长度长的序列。

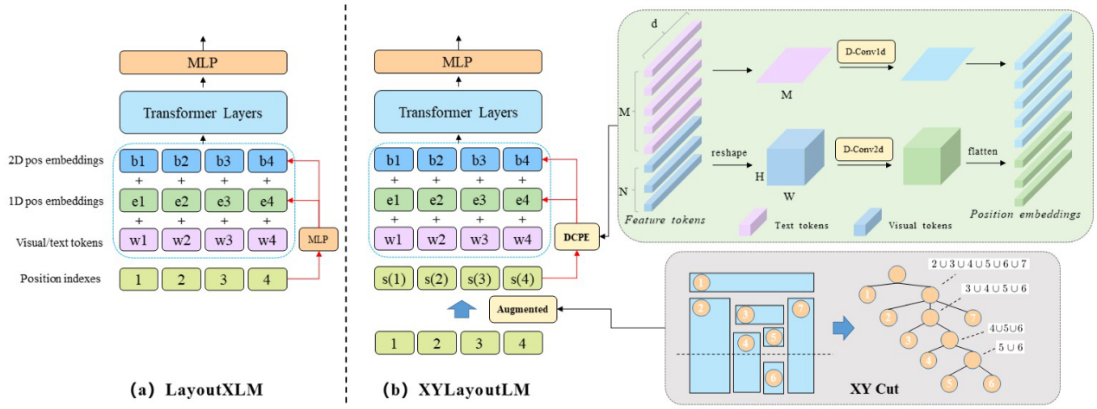


图 2.5 XYLayoutLM 的模型架构[38]。不同于 LayoutXLM，XYLayoutLM 提出了增强的 XY Cut 和 DCPE 来提取和利用布局信息，以实现多模态文档理解

早期的文档理解方法大多只关注布局或视觉特征，忽略了文档阅读顺序对模型学习的重要性。实际上，文档在读写过程中都有特定的顺序，这对模型的学习具有重要的参考价值。LayoutReader 是一个在 ReadingBank① 数据集上训练的多模态阅读顺序分析模型，能够有效提取文档的正确阅读顺序，但其模型较为复杂，推理时间较长，可能会影响文档理解的速度。XYLayoutLM 则提出了一个增强的 XYCut[41] 策略，通过对输入序列进行排列来获得合适的阅读顺序。此外，ERNIE-Layout[42] 也强调了文档阅读顺序对理解文档的重要性，并使用高级的文档布局分析工具对文本输入进行排序。与这些方法不同，GraphLayoutLM 依赖于图结构来对输入序列进行排序，这种方法被认为能更好地反映文本节点之间的逻辑关系。

GraphLayoutLM: 为了解决以上问题，Li 等人[43]提出了 GraphLayoutLM 模型，专注于文本结点之间的位置关系，并使用图数据结构来表现这种布局关系。GraphLayoutLM 在视觉富文档理解任务中展现出显著优势，其核心创新在于对文档布局结构的深度建模与动态优化。引入了图顺序优化和图 mask 优化两个优化策略。相较于传统方法，该模型基于图结构显式构建文本节点间的关系（如层级关联与空间邻接），通过图重排序算法动态调整阅读顺序，并结合图掩码策略优化自注意力机制，从而更精准地捕捉复杂布局中的逻辑关系。例如，在处理多栏文档或嵌套表格时，模型能够通过层级化重排序策略模拟人类阅读路径，有效解决了 LayoutReader 因依赖预训

① https://github.com/doc-analysis/ReadingBank

练模型而导致的推理延迟问题。此外，XYLayoutLM 等模型虽通过二维坐标编码反映位置信息，却忽视了文本间的语义关联，而 GraphLayoutLM 通过显式关系注入，在表单理解（FUNSD[44]）与多语言文档解析（XFUND[45]）任务中准确率提升约 15%。在效率方面，其轻量化图卷积架构较 LayoutReader 减少 30% 的参数量，同时支持实时处理低质量扫描文档，展现出更强的鲁棒性与实用性。

2.2.3 基于大语言模型的方法

（1）大语言模型

1）大语言模型的定义：大型语言模型（Large Language Model，简称 LLM）通常指的是基于 Transformer 架构的模型，其参数数量可达到数千亿甚至更多。这些模型通过对海量文本数据的训练，展现出卓越的语言理解与生成能力。例如，GPT-3[46]、PaLM[47]和 LLaMA[48]等均为典型的 LLM。大模型是人工智能领域中的一种深度学习模型，其核心特征在于通过对海量数据的预训练和迁移学习，能够高效地处理多种任务并具备良好的泛化能力。与传统的机器学习模型相比，大型模型通常拥有数亿至数万亿个参数，这使其在处理文本、图像等复杂非结构化数据时展现出显著的优势。

2）大语言模型的发展历程：大语言模型（LLMs）的演进始于 2017 年 Transformer 架构的提出，其自注意力机制对自然语言处理任务具有革命性影响，奠定了后续发展的技术基础（见图 2.6）。2018 年，早期大模型初露锋芒：Google 发布基于双向 Transformer 的 BERT 模型（3 亿参数），而 OpenAI 发布了单向序列建模的 GPT（1.17 亿参数），初步验证生成式模型的潜力。2019 年，模型规模加速扩张，OpenAI 发布 15 亿参数的 GPT-2，支持多任务学习；百度在同一时期发布了 ERNIE 系列模型（2019 年 4 月 ERNIE 1.0，11 月 ERNIE 2.0），通过知识增强技术优化语义理解能力[49-50]。

2020 年迎来规模突破，OpenAI 于 5 月发布 1,750 亿参数的 GPT-3，凭借 In-Context Learning 实现零样本推理；Google 推出 T5 模型[51]，统一 NLP 任务框架。2021-2022 年，多任务与代码生成成为焦点：OpenAI 基于 GPT-3 开发代码生成模型 Codex[52]（2021 年），Google 发布 5,400 亿参数的 PLaM 模型（2022 年 4 月），探索超大规模预训练的边界。同期，Meta 开源 650 亿参数的 LLaMA 模型[48]，推动开源生态发展。

2023 年进入多模态时代，OpenAI 于 3 月推出支持文本、图像的多模态 GPT-4（参数达 100 万亿）[53]，Google 发布 5,620 亿参数的 PLaM-E[54]，实现视觉-语言深度

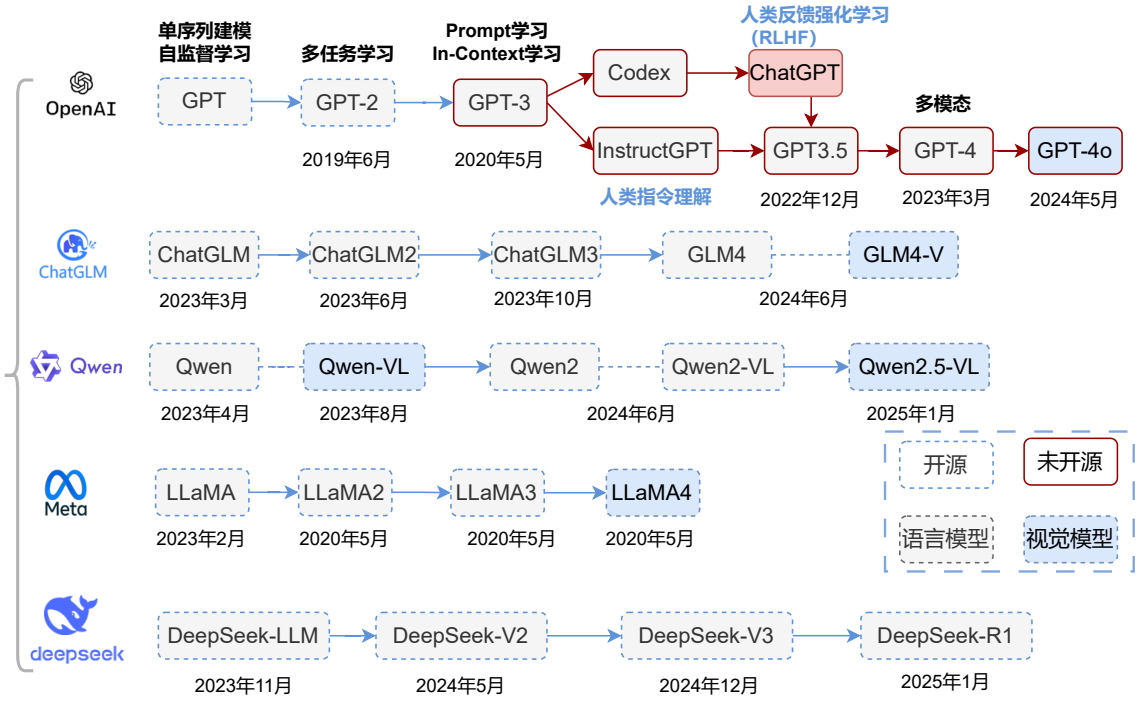


图 2.6 大语言模型的发展历程

融合。百度发布 70-650 亿参数的文心一言（ERNIE 3.0），Meta 升级 LLaMA 至 690 亿参数的 LLaMA2[55]（2023 年 7 月）。开源模型 LLaMA3[56]（706 亿参数）于 2023 年 7 月首次在基准测试中击败闭源模型 Gemini[57] 与 Claude，标志开源生态的崛起。

2024 年，推理与架构创新主导发展：OpenAI 发布了 GPT-4.5（2024 年 9 月），这是一种复杂的增强逻辑，谷歌发布了基于混合专家架构的 Gemini 1.5[58]（2024 年 2 月）和 Gemini 2.0（2024 年 12 月）。智谱发布了 GLM-4 系列模型[59]，将包含对话、推理和反思等多种能力的模型提升到 32B。阿里团队团队推出的 Qwen2.5-VL 模型[60]，进一步增强了视觉能力，持续优化其在视觉相关任务中的表现。DeepSeek 系列迭代加速，2024 年 5 月发布多模态 DeepSeek-V3[61]，12 月发布 DeepSeek-R1[62] 深度推理模型，模拟人类推理进行高级逻辑推理。

总体来说，LLMS 从单向建模阶段开始，经过规模化、多任务学习、token 生成、多模型融合，在来回推理和动态推理方面逐渐发展，不断拓展自然语言处理、跨模型理解、通用人工智能的边界。

3）几种常用的大语言模型的比较：LLAMA3[56] 是一个基于 Transformer 框架的多模态模型，其中多模态数据针对文本、图像和布局进行了优化。其主要组件包括文本编码器、图像编码器、编码器和多模型融合单元。文本编码器使用 TransFormer

模型处理文件中包含的文本信息，捕捉含义和上下文之间的关系。图像编码器通过卷积神经网络或视觉 Transformer 从图形、表格和插图等文件中提取图像特征。编码器处理与文档布局相关的信息，例如文本位置、字体大小、段落结构等，这有助于模型理解文件的物理结构。多模型融合模块通过注意力机制和跨模态交互层深度整合文本、图像、图形信息，生成统一的数据。LLAMA3 具有多模型融合、集成能力强、端到端学习、有效问答能力等优势。它能够深度整合文本、图像和图形信息，全面理解文档内容。通过大量的初步训练，LLAMA3 能够处理多样化的文档类型，完成复杂的文档理解任务。

GPT-3.5 是 OpenAI 开发的一种生成式大型语言模型，在自然语言处理（NLP）领域表现出色。其主要优势包括强大的语言生成能力、广泛的知识覆盖、多任务处理能力和高效的上下文理解。GPT-3.5 能够生成流畅、连贯且语义丰富的文本，适用于文本补全、对话生成和内容创作等任务。通过大规模预训练任务，GPT-3.5 学习了各个领域的知识，能够回答各种的问题，而且能够提供解答的细节。与此同时，GPT-3.5 还能够理解长文本的上下文关系，并根据上下文生成准确的回复，这使得它在对话系统和文档理解任务中表现优异。

GPT-4.0 是 GPT-3-5 的加强版，建模能力更强，准确率更高，熟练度更高，上下文理解能力更强，推理能力更强。GPT-4.0 不仅支持文本处理，还可以处理图像、表格等数据，通过结合视觉和文本信息，更全面地理解复杂文档。在语言生成和问答任务上，GPT-4.0 的准确率有很大的提升，能够给出更准确的答案，减少错误和偏差。GPT-4.0 能够处理更长的上下文，理解更复杂的语义关系，提高了处理长文件和多轮对话状态的性能。

GPT-3.5 和 GPT-4.0 在文档理解任务中均表现出色，但各有侧重。GPT-3-5 在文本解析、信息提取和文档分类任务上表现出色，可以从文档中提取关键信息并处理结构化文本。然而，GPT-3-5 严重依赖文本信息，对视觉丰富的文档的理解能力有限。GPT-4.0 在视觉文档理解和复杂文档处理任务上表现更佳。它不仅能从文档中提取关键信息，还能处理更复杂的文档结构，并从多种建模数据（如图像和表格）中提取信息。

（2）视觉语言大模型

传统的视觉文档领域初始训练模型需要在数据预处理阶段使用现成的 ORR 引擎获取文档图像的文本和图像信息。这不仅大大增加了文档图像预处理的成本，而

且还将文本识别过程与模型训练过程分离。为了解决上述问题，适用于视觉语言理解任务的可视化大型语言模型（VLLMS）由于 OCR 引擎的质量问题以及无法公平地执行真实模型，逐渐被排除在 OCR 引擎的限制之外。大规模视觉语言模型在 VRDU 领域的应用主要体现在对文档内容的深度理解和分析。这些模型结合了大型语言模型和光学编码器，通过结合计算机视觉（CV）和自然语言处理（NLP）技术，使机器能够通过视觉建模和文本来可视化世界并提取其中的意义。它通常使用从网络上收集的数百万张图像和文本，并提供具有泛化和迁移能力的表示。所以只需通过简单的自然语言描述和提示，这些预训练的基础模型就可以被应用到下游任务。

GLM-4V[59]是智谱 AI 推出的基于 GLM-4 系列的多模态大模型，专注于图像识别和多模型交互。该模型建立在自然语言处理领域丰富的 GLM 模型基础之上，并在此基础上扩展了视觉处理能力，将文本与图像深度融合。GLM-4V 能够通过图像和文本信息的深度融合，充分理解多媒体内容，显著提高识别的准确性和可行性。它支持端到端学习，直接从图像和文本生成结构化输出，简化处理并提高效率。此外，通过大量的前期训练，GLM-4V 系统能够处理多样化的图像和文本类型，具有广泛的泛化能力，适用于多种应用场景。

Qwen2-VL[63]是由阿里通义千问团队开发的 Qwen 模型系列[64-65]中最新版本的视觉语言模型, 如图 2.7所示。该模型在多个视觉理解基准测试中实现了最先进的性能，包括 MathVista[66]、DocVQA[67]和 MTVQA[68]等。和上一代模型相比，Qwen2-VL 在结构和功能方面实现了创新，大幅度提高了处理和推理能力。例如，Qwen2-VL 提供对原生动态分辨率的全面支持，能够处理任意分辨率的图像输入。至少有 4 张不同大小的图像被转换为动态数量的图像到的 Token。这样的设计不仅保持了输入与原始图像信息的一致性，而且模仿了人类视觉的自然感知，增加了模型在图像处理任务中的灵活性和效率。除了以上创新，Qwen2-VL 还能够通过图像和文本信息的深度融合，充分理解多媒体内容，大大提高识别的准确性和可行性。结构上，Qwen2-VL 采用了多模态旋转位置嵌入。传统的旋转位置包含仅捕获 1D 序列的位置信息，而多模态旋转位置嵌入通过将旋转分为时间、高度和宽度维度，可以同时捕获 1D 文本、2D 图像和 3D 视频的位置信息。这一创新增强了模型对复杂多模型数据的理解和建模能力，为其在多模型任务中的发挥提供了重要支持。Qwen2-VL 模型支持从头到尾的学习方法，直接从图像和文本生成结构化输出，简化处理并提高效率。通过大量的初步训练，该模型可以处理多种图像和文本类型，提供广泛的全局能力，可应用

于各种应用场景。

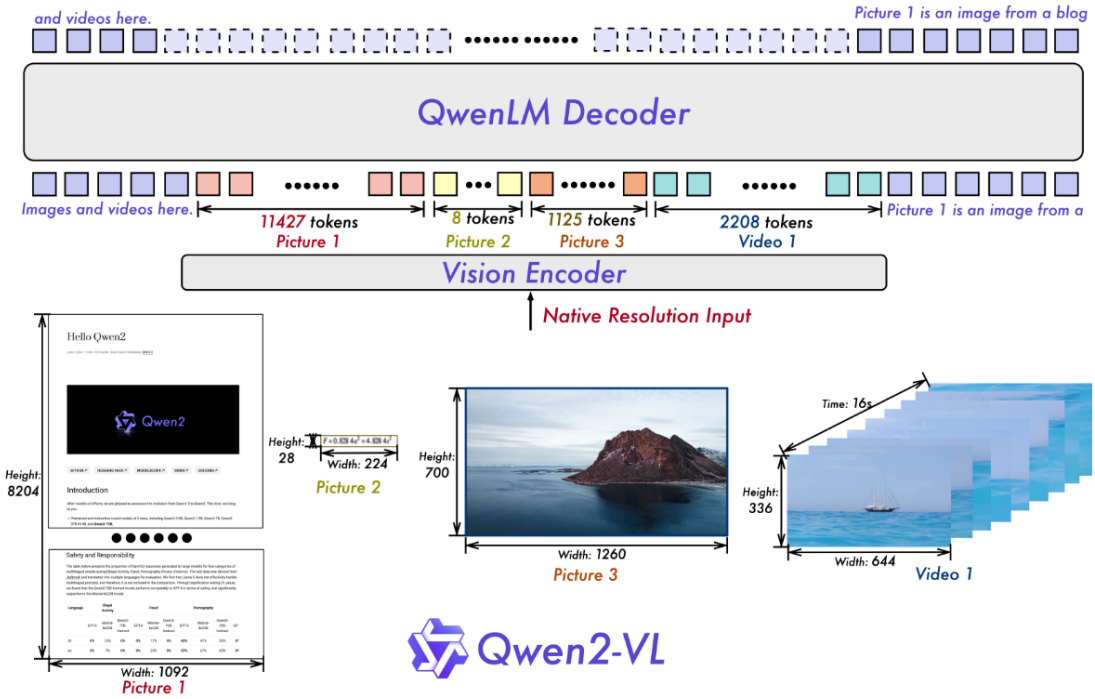


图 2.7　Qwen2-VL 模型架构[63]

2.3　人类认知特征研究

认知是获取、储存、传递和利用信息的过程[]10.1177/00144029211050860，大脑对信息的处理是认知活动的基础。视觉是人们获取信息的重要渠道，包括对信息的感知、储存、记忆和解码等一系列心理过程。大约 80% 到 90% 的信息来自视觉。所以视觉是研究人类认知的重要途径。近年来，虽然各种认知生理测量技术迅速发展，由于眼动追踪测量方法准确、操作方便、成本低廉，在认知研究领域得到了广泛的应用。认知心理学研究发现，人类阅读过程中有两个主要的认知因素：阅读顺序和视觉注意。这些认知过程使读者能够专注于文档的重要区域，优先处理相关信息，并理解复杂的视觉环境。人类的阅读顺序和注意机制是人类认知特性中的重要部分，它们共同影响着阅读的效率和质量。

阅读顺序通常受语言习惯和视觉引导的影响。在大多数语言中，阅读方向是从左到右、从上到下。在阅读过程中，眼睛通过快速扫视和短暂注视获取信息，重要内容往往获得较长的注视时间。此外，上下文也会对阅读顺序产生影响。读者会根据上下文预测后续内容，并可能跳过不重要的部分。

同时，注意力机制在阅读过程中发挥着重要作用。首先，人类具有选择性注意的能力，能够首先关注重要的信息，忽略不相关的内容，例如标题和关键词更容易引起注意[69-70]。其次，注意力的持续时间有限，通常为 20 至 30 分钟。此后效率就会下降，需要适当休息。此外，读者还会根据任务要求分配注意力。例如，浏览时注重关键词，深入阅读时注重细节。在阅读过程中，注意力也可以在不同的信息之间转换，比如从文本切换到图表。最后，外界干扰（如噪音）会分散注意力，从而降低阅读效率。

眼动追踪技术用于记录眼睛在视觉刺激物（如文本、图像、视频等）上的运动情况，能够提供与明显视觉注意相关的信息，包括注意的对象、注视时长、注视顺序等。[71] 综上所述，通过眼动追踪技术可以得到人类阅读的顺序以及注意力的分布情况，进而研究人类的认知机制。下面，我们将从眼动追踪技术、人类阅读的顺序、人类注意力的机制三个方面进行阐述。

2.3.1 眼动追踪技术

眼动追踪是一种通过记录眼球运动来研究视觉注意力、信息处理和认知过程的方法。该技术可以精确测量个体在进行观察、阅读等视觉任务时眼睛的位置、移动速度和停留时间，从而深入了解个体的认知状态和过程。该技术可以精确测量个体在观察、阅读等视觉活动时眼球的方向、移动速度和停留时间，从而深刻揭示个体的认知状态和过程。

眼动仪作为眼动追踪技术的应用工具，其工作原理是利用红外光照射人眼，引发瞳孔与角膜之间的反射现象。然后，通过摄像机精确测量两者之间的距离，以确定眼球的注视点。其核心技术在于精确捕捉和分析眼球的细微运动。眼动仪主要分为三种类型的设备：台式眼动仪、头戴式眼动仪和嵌入式设备[72] 台式眼动仪适用于实验室环境，对被试的位置有固定的要求。头戴式眼动仪便于携带，适合在自然环境中监测。嵌入式设备融入到日常生活用品中，方便在使用过程中进行眼球运动追踪。眼动追踪有多种指标，例如注视位置、时间、路径、频率和分布等。热力图可以直观地显示关注的区域。关注区域（AOI）用于特定区域的注视分析，同时还记录重访、首次注视时间和停留时间等。眼动大致可分为注视和眼动两类。注视，即眼球停留在某一位置，是获取和处理视觉信息的关键时期。而眼跳则是眼球快速跳转到另一个位置的瞬间，常用于在复杂的视觉环境中切换注意力焦点。通过测量眼动参数，

研究人员可以深入了解个体在执行特定任务时的注意力分配、信息处理方式以及认知负担状况。[73] 眼动追踪技术在阅读行为分析、心理与认知科学研究、人机交互优化、虚拟现实等多个领域有着广泛的应用。眼动追踪是一种通过记录眼球运动来研

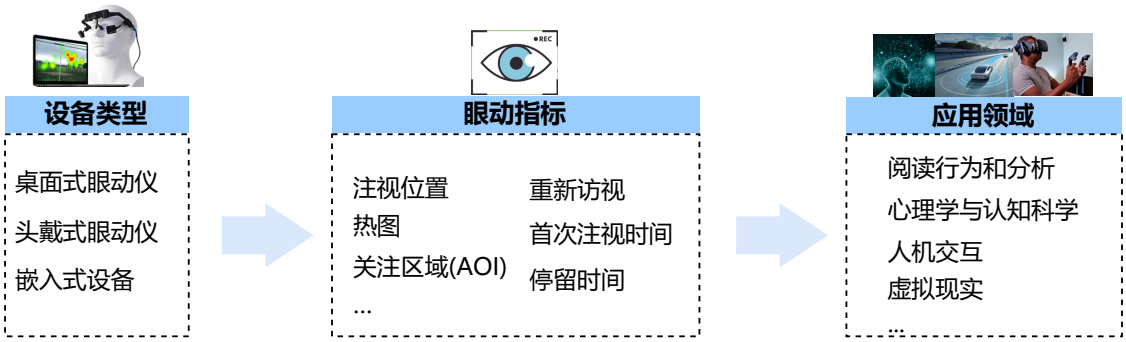


图 2.8　眼动追踪技术

究视觉注意力、信息处理和认知过程的方法。该技术可以精确测量个体在进行观察、阅读等视觉任务时眼睛的位置、移动速度和停留时间，从而深入了解个体的认知状态和过程。该技术可以精确测量个体在观察、阅读等视觉活动时眼球的方向、移动速度和停留时间，从而深刻揭示个体的认知状态和过程。

眼动仪作为眼动追踪技术的应用工具，其工作原理是利用红外光照射人眼，引发瞳孔与角膜之间的反射现象。然后，通过摄像机精确测量两者之间的距离，以确定眼球的注视点。其核心技术在于精确捕捉和分析眼球的细微运动。眼动仪主要分为三种类型的设备：台式眼动仪、头戴式眼动仪和嵌入式设备[72] 台式眼动仪适用于实验室环境，对被试的位置有固定的要求。头戴式眼动仪便于携带，适合在自然环境中监测。嵌入式设备融入到日常生活用品中，方便在使用过程中进行眼球运动追踪。眼动追踪有多种指标，例如注视位置、时间、路径、频率和分布等。热力图可以直观地显示关注的区域。关注区域（AOI）用于特定区域的注视分析，同时还记录重访、首次注视时间和停留时间等。眼动大致可分为注视和眼动两类。注视，即眼球停留在某一位置，是获取和处理视觉信息的关键时期。而眼跳则是眼球快速跳转到另一个位置的瞬间，常用于在复杂的视觉环境中切换注意力焦点。通过测量眼动参数，研究人员可以深入了解个体在执行特定任务时的注意力分配、信息处理方式以及认知负担状况。[73] 眼动追踪技术在阅读行为分析、心理与认知科学研究、人机交互优化、虚拟现实等多个领域有着广泛的应用。

2.3.2 人类阅读顺序

阅读是一个复杂的认知过程。阅读认知的主要外部指标是认知效率和认知效果，代表读者花费的时间和精力，以及是否能对文章作出正确的判断[74]。认知策略可反映认知的内在机制，即人们独立控制自己内在心理活动以获得新知识的方式。视觉是获取信息的重要渠道，眼动是视觉的直接反映，能为分析认知过程提供准确而丰富的信息[75]，通过眼动追踪考查阅读中的认知加工过程。阅读顺序对视觉富文档理解具有重要影响。人类阅读顺序是指人们在阅读文本时，视线扫描的顺序和方向。在通常情况下，人们会按照从左到右、从上到下的顺序进行阅读，以此获取信息并理解文本。其形成与语言文字的书写方式相关。在大部分语言文字中，包括中文、英文等，文字的书写方式都是从左到右、从上到下的。因此，人们在阅读时也习惯性地按照这个顺序进行视线扫描。

（1）眼动扫描路径

人类眼球主要通过（Fixation）、眼跳（Saccade）和扫描路径（Scanpath）三种基本参数来表示眼动阅读顺序可视化的过程。注视时，眼睛的中央凹对准目标，视线看似停留，实则伴随着微小的漂移、震颤和不随意眼跳。眼跳是视线在不同注视点间快速跳跃的过程，这种跳跃速度极快，让人误以为视线是平滑移动的。追随运动则分为两种：一种是头部固定时，眼睛追随运动物体；另一种是头部或身体运动时，眼睛进行补偿性眼动以保持物体在中央凹上。这些运动模式确保了人类能够高效地捕捉和处理视觉信息。为了量化眼球运动，研究人员定义了几个关键统计指标。注视顺序反映了兴趣变化和对不同区域关注度的转移。注视持续时长则衡量了信息提取的难易程度和目标的吸引力，时长长的表明信息复杂或目标吸引力强。兴趣区内注视次数是区域重要性的指标，次数越多，区域对观察者越重要。

如图 2.9所示，眼动扫描路径可以进行形式化定义：注视点定义为 $F=(x,y,t)$，其中 (x,y) 表示注视点的坐标，t 表示该注视点的持续时间。在目标分析区域预先设定感兴趣区（Area of Interest, AOI），可统计该 AOI 内的注视点数量。若已知 AOI 内所有注视点的持续时间总和，则可计算注视点平均持续时间，即总持续时间与注视点数量之比。眼跳定义为 $S_i=(F_{i-1},F_i)$，其中 F_{i-1} 和 F_i 为相邻的两个注视点，眼跳即为从 F_{i-1} 指向 F_i 的向量。眼跳数目即为某段时间内所有眼跳的个数。扫描路径定义为 $P=(S_1,S_2,\ldots,S_n)$，其中 S_i（$i=1,2,\ldots,n$）为眼跳，扫描路径是多个连续

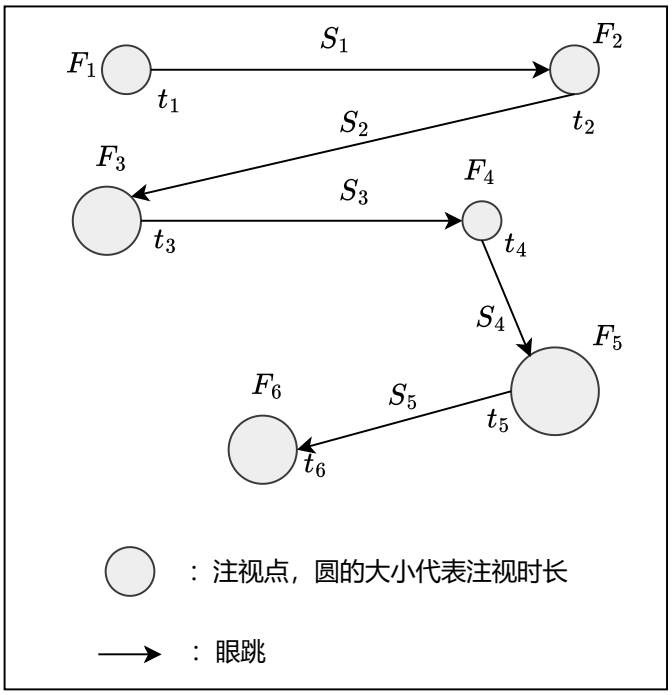


图 2.9　扫描路径

眼跳构成的序列集合。其中扫描路径长度是路径中各眼跳长度的总和；扫描路径时长是整个扫描路径消耗的时间总和，包括所有注视点的持续时间和眼跳时间。

（2）眼动阅读模式

尽管文本的阅读顺序一般遵循相对固定的模式，但视觉丰富的文件通常以混合模式（如文本、图像、图形等）呈现信息，具有复杂的二维布局和语义结构[37,76]。读者通常通过考虑文本内容的空间结构、图像信息和文件的排版位置来综合选择合适的阅读顺序[77]。这种自然的时间处理可以帮助读者更好地理解文件的内容和意图，并建立不同模式之间的联系。例如，当看到一张图时，读者通常会先从视觉上分析它的整体，对它所描述的内容有一个大致的了解。从那里，他们通常首先关注最明显的数据或标签，然后逐渐扫描其余的数据和标签，以获得对所呈现信息的更完整的理解。

人类在阅读视觉富文档时一般有四种四种阅读模式[78]，如图 2.10所示：

- **Normal-Z**：也称为“Z 字形模式”，指的是人们在阅读纯文本或弱结构的文档时出现的典型眼动阅读模式。这种“Z 字形模式”是人类视觉系统处理和理解文本的有效方式。具体来说，在阅读过程中，我们的眼睛会沿着一行文字从左到右移动，然后跳回到下一行的开头，形成一个之字形的眼动扫描轨迹。如图

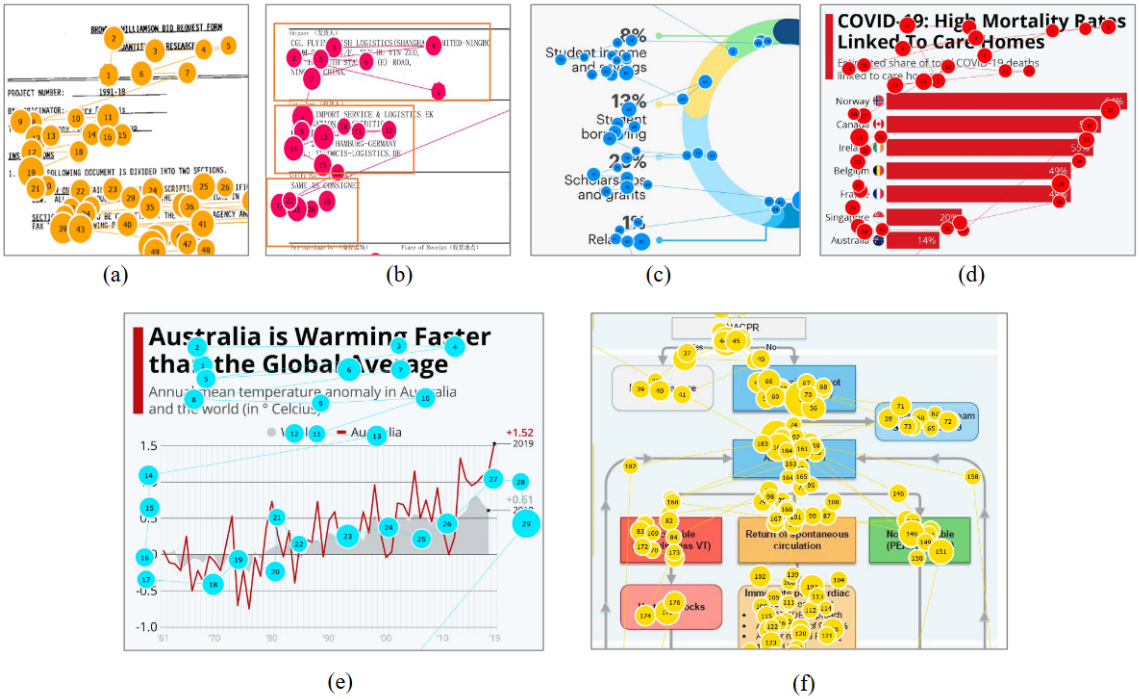


图 2.10 显示人类阅读时眼球运动模式的散点图可以用多种方式组织：(a) 正常-Z 顺序；(b) 局部优先顺序；(c, d, e) 跨模态交互顺序；(f) 视觉引导[78]

所示，显示了人类阅读时眼球运动序列的示意图，它通常类似于字母“Z”的形状。

- **局部优先**：局部优先的眼动模式是人类在阅读时处理表格或表格的分层布局结构时常用的一种认知策略。通常，如图 2.10（b）所示，通过局部信息优先，我们会首先关注表格单元格内的内容，然后在表格或表单文档中查找时将注意力转移到单元格周围的其他元素上。这有利于快速阅读和检索信息文本，获取所需信息。
- **跨模态交互**：当人们阅读图表、条形图或其他形式的图形时，他们往往会表现出一种反应性的表面眼球活动模式。阅读时，眼睛沿着图表和相关文本之间的路径移动。例如，当人们阅读地图时，他们通常首先关注图表的各个部分，向外寻找相应的文本标签，眼睛来回移动，从而形成多模式交互模式。这种模式源于地图独特的形状和结构，地图通常包含最重要的信息。相比之下，文字丰富的部分提供了详细的信息，人们需要逐步浏览才能理解。因此，这些类型的信息图中的眼动追踪揭示了一种趋势模式。
- **视觉引导**：阅读流程图时，经常会出现回顾性眼动模式，因为读者必须理解和比较框架中的不同文本。流程图通常包含大量信息和细节，并以有组织、有逻

辑的方式呈现。读者通常首先关注地图的总体结构和主题，然后再关注所有细节。当需要参考上下文文本时，读者可以重读前面的章节，以帮助更好地理解信息的含义。这种纠正和反馈模型反映了读者处理信息的能力，有助于提高对信息的理解和回忆。

2.3.3 人类注意力机制

注意力机制是心理学、神经科学和人工智能领域的一个重要概念，通过该机制可以调配人类的认知资源来获取相关信息。它是人类智能系统的基本功能之一，在认知过程中发挥着重要作用[79]。Wolfe[80]将注意力定义为“从众多无关的刺激、反应、记忆或思维中筛选出与当前行为相关的内容的心理能力”。这个概念揭示了信息加工的选择性和方向性。传统模型，例如有限资源理论和瓶颈模型，强调了注意力在优先考虑内容以防止超载方面的作用。在阅读过程中，焦点会根据阅读的内容和上下文而变化，当同时处理一组任务时，会对阅读表现产生负面影响。最近的一项研究中使用的理解视觉丰富文件（VRD）的 AI 模型利用眼动追踪数据和视觉搜索理论来模拟人类的注意力，重点关注标题和数字等关键领域[81]。通过在批量认知集成中使用 EMA-T 模型，AI 模型可以更好地验证这些领域的优先级，并模拟人类在排列复杂文件时的注意力[82]。

研究表明[83]，人类视觉系统 (Human Visual System) 具有极强的处理感官数据的能力，对于人类来说，其处理速度每秒约有 $10^8 \sim 10^9$ 字节的数据。认知科学研究认为[84]，视觉聚焦机制是人类视觉系统处理海量数据能力的基础。在对视觉数据进行视觉处理的初始阶段，人类视觉系统能够快速聚焦于场景中的重要区域。这种选择性感知大大减少了人类视觉系统需要处理的数据量，使得个体在面对复杂的视觉信息时能够抑制不相关的刺激，将有限的神经计算资源分配给场景的关键部分。这个过程为更高层次的认知推理提供了更容易、更重要的信息。

从生理机制的角度，巴甫洛夫通过定向反射理论对注意力的产生进行了深入探讨。他指出，注意力的生理基础源于大脑中枢神经系统中兴奋与抑制过程的相互作用和相互诱导[85]。这种动态平衡使得个体能够对外界刺激做出选择性反应。

根据生理学的研究，注意力通常被分为两种类型：自下而上（bottom-up）和自上而下（top-down）[86]。自下而上的注意力机制主要受到外部刺激和特征的驱动，其特点是快速、自动且无意识，能够迅速捕捉环境中的显著变化并调整注意力焦点。相

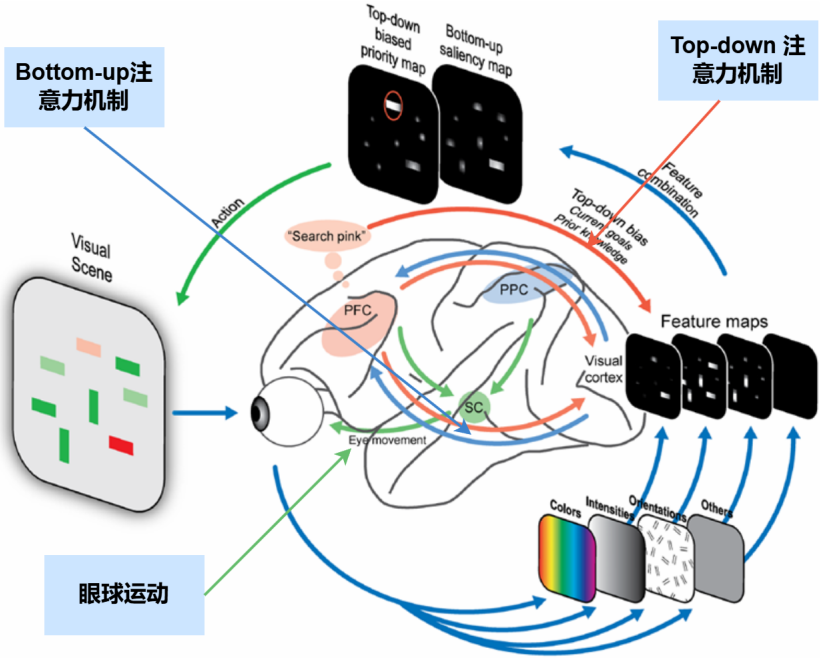


图 2.11 眼球运动过程中产生的自下而上和自上而下的注意力过程的示意图[86]

对而言，自上而下的注意力机制则是由任务目标和个人经验（包括记忆）所驱动，表现出更强的主动性和目的性[87]。

Bottom-up 注意力机制: 自下而上的注意力机制是一种由外部刺激驱动的、迅速且自动化的注意力过程。该机制主要依赖于环境中显著特征或变化的存在，例如突发的声音、鲜艳的颜色或强烈的对比度。在视觉系统中，自下而上的注意力与视网膜及大脑的上丘区域密切相关。

自下而上的注意力机制主要依赖于大脑的早期视觉处理区域，包括视网膜、外侧膝状体（LGN）和初级视觉皮层（V1）。这些区域负责检测环境中的显著特征，并将相关信息传递至更高级的视觉处理区域。上丘在这一过程中发挥着重要作用，负责快速定位和响应环境中的显著刺激。上丘在大脑中具有重要地位，尤其是在自下而上的注意力机制中。它接收来自视网膜的低分辨率、去饱和的视觉信息，并生成视觉显著性图[88]。该显著性图能够识别环境中最为显著的区域，从而迅速引导注意力转向这些区域。此外，上丘还直接控制眼部肌肉，触发反射性扫视运动，使眼球能够迅速将高分辨率的中央凹对准显著区域，以获取更为详细的视觉信息。

自下而上的注意力机制是一种特征驱动的过程，其特点在于注意力被环境中显著的特征所吸引。这些特征包括颜色、亮度、运动和方向等。由于该机制不依赖于高级认知过程，而是直接受到外部刺激的驱动，因此能够在极短的时间内（通常为

毫秒级别）实现注意力的转移与集中。此过程无需有意识的控制或努力，而是由外部刺激自动引发。这种自动化的机制使个体能够在复杂环境中迅速筛选出重要信息，而不必消耗大量的认知资源。

通过自下而上的注意力机制，个体能够在复杂环境中快速识别并响应重要刺激，从而提升视觉处理的效率。这一自动化的注意力过程不依赖于个体的主观意图，而是基于外部刺激的显著性以及大脑的快速反应机制。

Top-down 注意力机制: 自上而下的注意力机制主要依赖于个体的目标、意图及其既有经验。这一机制能够根据任务的需求主动分配注意力资源，体现出一种主动且有计划的注意过程。与自下而上的注意力机制不同，自上而下的注意力过程是由任务目标、意图和个体经验所驱动的，具有主动性和意识性，并依赖于个体的认知控制及目标导向行为。

在自上而下的注意力机制中，大脑的高级认知区域（例如前额叶皮层）会根据当前的任务需求或目标，主动引导注意力资源向特定的刺激或信息进行分配[89]。这一机制使得个体能够忽略无关的刺激，专注于与任务相关的内容。例如，当个体在嘈杂的环境中寻找特定的声音时，即便周围存在更多显著的刺激，仍然能够将注意力集中于目标声音上。

自上而下的注意力机制涉及以下几个关键过程：

1) **目标设定**：个体根据当前的任务或目标，确定需要注意的信息类型或区域。
2) **注意力分配**：大脑的高级认知区域根据目标，主动调控感觉器官（如眼睛）和感知系统，将注意力资源分配到相关刺激上[90]。
3) **信息筛选**：在自上而下的控制下，个体能够过滤掉无关的信息，专注于与任务相关的内容[91]。
4) **反馈调节**：根据任务进展和环境变化，个体可以动态调整注意力的分配，以优化任务执行。

自上而下的注意力机制在复杂任务的执行及目标导向行为中具有重要意义。该机制使个体能够在动态环境中，根据内在目标和意图灵活调配注意力资源，从而提升认知效率和任务表现。这一机制反映了人类认知系统的高度适应性与主动性。

与自上而下的注意力机制相对的是自下而上的注意力机制。自上而下的注意力是由特定目标或任务驱动的，依赖于高级认知过程，例如前额叶皮层的调控。而自下而上的注意力则是由外部刺激所驱动，依赖于早期视觉处理区域的快速反应。例

如，在驾驶过程中，突如其来的障碍物或交通信号灯的变化会迅速吸引驾驶员的注意力，从而有效避免潜在的危险。

第三章　基于眼动追踪与预排序模型的仿人类阅读顺序嵌入方法

3.1　引言

视觉富文档理解（VRUD）任务旨在处理各种类型的半结构化文档，如发票、收据和表单。该任务通过将 OCR 识别的序列化结果输入多模态模型中进行标注，由于文档的多样性和复杂性，OCR 识别后的顺序往往是乱序的，不利于文档智能模型在下游任务中的效果，因此序列化后的顺序成为了视觉富文档理解的关键。单纯的文本信息不足以从不同的文档类型中提取关键信息，因此需要一种多模态的方法，通过联合建模考虑多种模式的一致性和相关性，包括文本、视觉和布局。

近年来，文档智能领域的研究取得了显著进展，尤其是在多模态信息融合技术的推动下，研究者们通过整合文档中的文本、布局和视觉等多种模态信息，成功提取了文档中的结构化知识。现有的模型如 LayoutLM 和 StrucText 系列[5-10,40]虽然这些模型可以获得细粒度的多模态文档表征，并在下游视觉富文档（VRD）理解任务中取得显著效果，但缺乏产生适合 Transformer 架构的序列化输入顺序的能力。

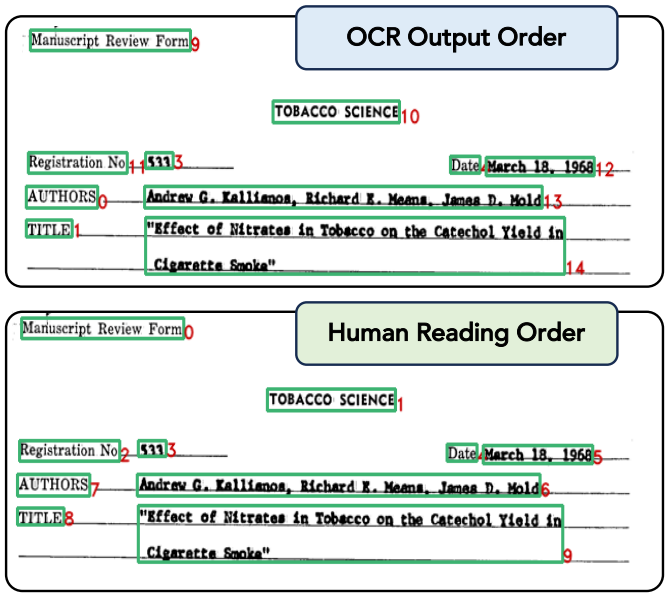


图 3.1　OCR 引擎输出的序列化输入顺序与实际人类阅读顺序的案例比较，红色数字表示阅读顺序中的序号[78]

而在认知心理学中，人类在阅读时对不同的文档结构会产生不同的顺序以适应

人类对文档的理解。基于以上问题，我们通过眼动追踪实验采集了人类的阅读顺序，推出了 DocTrack 数据集，这是一套包含人类眼动信息的基准数据集，旨在探索如何将人类的阅读顺序数据融入现代文档 AI 模型中。我们设计了一个预处理流程，将输入序列依据人类的视觉浏览顺序重新排序，在语义实体识别和文档问答两个子任务上进行了实验，初步验证了人类阅读顺序对视觉富文档理解任务有所提升。

本章的三个主要贡献包括：构建 DocTrack 数据集，研究类人阅读顺序生成方法，以及提供对人类阅读视觉丰富文档的更全面理解。通过这些贡献，我们旨在为未来开发更先进的文档 AI 模型及模仿人类阅读习惯的机器提供支持。

3.2 问题定义

本文主要关注视觉富文档理解任务中语义实体识别和文档文档两个子任务，定义如下：

语义实体识别（Semantic Entity Recognition, SER）是信息抽取（Key Information Extraction, KIE）中的一项重要任务，它涉及对文档图像中的每一个检测到的文本进行分类，是一个序列标注任务，例如将标注为姓名、身份证等类别。数学定义如下：给定一个由扫描图像 I 和 OCR 边界框内文本段的列表 $B = \{b_1, b_2, \ldots, b_N\}$ 组成的文档 D，我们的目标是找到一个函数 F_{SER}，它能够预测文档 D 中每个标记的相应实体类型。预测的标签序列 $\mathcal{E}$ 使用了“BIO”标注方案 $\{Begin, Inside, Other\}$ 和一个预定义的标签集。

形式上，我们可以将 SER 任务定义为寻找一个最优的标签序列 $\mathbf{y} = \langle y_1, y_2, \ldots, y_N \rangle$，其中每个 $y_i \in \mathcal{E}$ 并且 y_i 表示文本段 b_i 的预测实体类型。目标是最大化给定输入 D 的条件概率 $P(\mathbf{y}|D)$。

$$F_{\text{SER}} : B \times I \rightarrow \mathcal{E} \tag{3.1}$$

其中，$\mathcal{E}$ 表示所有可能的实体类型组成的集合，而 F_{SER} 是一个从文档 D 到实体类型集合 $\mathcal{E}$ 的映射函数。预测的标签序列 $\mathcal{E}$ 使用了“BIO”标注方案，其中：B（Begin）表示实体的第一个标记，I（Inside）表示实体内部的标记（除第一个标记外），O（Other）表示非实体的标记。

文档问答（Documents Question Answering，DQA）是一项结合计算机视觉（CV）和自然语言处理（NLP）的人工智能任务。其目标是让计算机根据给定的图像和相关的自然语言问题，生成准确的自然语言答案。具体来说，DQA 任务模型需要理解图像内容、问题的语义，并结合常识知识来推理出正确的答案。给定一个文档 D 和一个与文档相关的问题 Q，DQA 任务的目标是找到一个答案 A，使得 A 能够正确回答 Q 关于 D 的询问。这个过程可以形式化为：

$$A = \arg\max_{a \in \mathcal{A}} P(a \mid D, Q) \tag{3.2}$$

其中，$\mathcal{A}$ 是所有可能答案的集合，$P(a \mid D, Q)$ 是在给定图像 D 和问题 Q 的条件下，答案 a 的概率。

3.3 模型架构

为了更贴近人类的阅读习惯，我们引入了一种称为“预排序”（作为预处理的重排序）的方法。这一概念源自统计机器翻译领域，其核心思想是在处理输入数据之前，先按照人类自然阅读的顺序对这些数据进行重新排列[92-94]。通过这种方式，我们能够更直观地展示如何将多模态序列输入特征与人类的阅读顺序对齐。这种对齐不仅有助于模型更好地理解输入信息，还能让我们更清晰地评估阅读顺序对视觉丰富文档（VRD）理解任务的影响。

具体而言，预排序过程模拟了人类在阅读复杂文档时的自然行为。例如，当面对包含文本、图像和布局信息的多模态文档时，人类通常会按照一定的顺序（如从左到右、从上到下）依次处理这些信息。通过预排序，我们可以将输入特征重新组织为这种自然的阅读顺序，从而为后续的处理步骤提供更符合人类认知习惯的输入数据。这种方法不仅提高了模型对输入数据的理解能力，还为研究阅读顺序对任务性能的影响提供了一个有效的实验框架。

如图 3.2所示，我们提出的实验流程，首先我们会使用初始的 OCR 顺序输入到预排序模型中，预排序模型通过参考人类阅读顺序，负责比较每张文档中两个边界框的的前后位置，然后再通过排序算法将原始顺序重新排序，最后将按照生成的顺序输入到下游任务中。这样我们不仅考虑人类阅读顺序，不同的排序模型也会根据

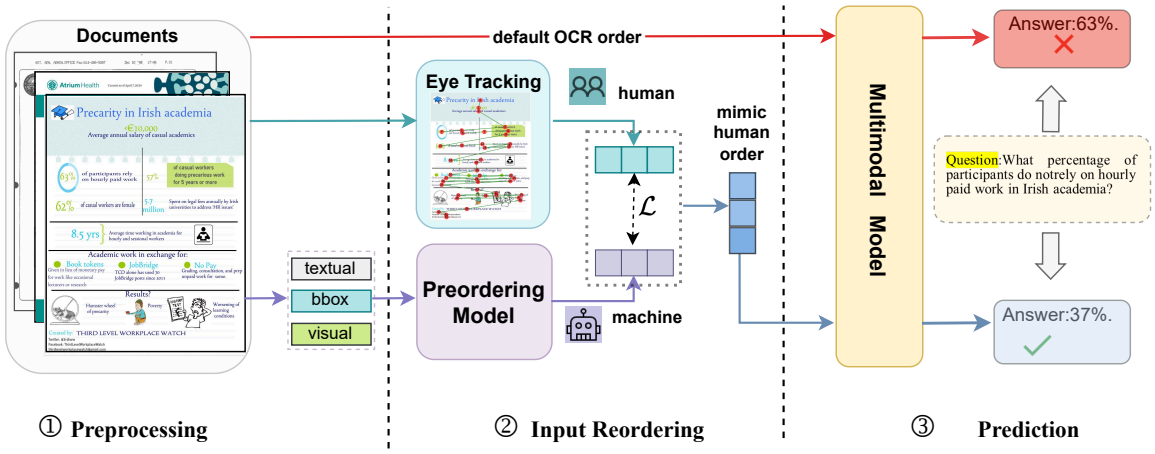


图 3.2　方法架构图

特征对两个边界框中的特征进行排序。

3.3.1　人类阅读眼动顺序获取

（1）眼动追踪实验

1）眼动实验设计与数据采集。为了开展眼动实验，我们首先将数据集随机划分为五个部分，并招募了五名在校研究生或本科生作为实验参与者。实验流程如图 3.3所示，实验设备包括一台 24 英寸、分辨率为 1080P 的惠普显示器，以及 Tobii TX300 眼动仪和 Tobii Studio 软件。实验过程中，参与者被要求在显示器上阅读指定的数据内容，同时眼动仪实时记录其视觉运动轨迹。Tobii TX300 眼动仪能够以高精度捕捉参与者的眼动行为，并将其转化为数字化的眼动数据。具体记录的数据指标包括注视点位置、注视持续时间、注视频率、眼跳距离以及瞳孔大小等。这些数据为后续分析参与者的认知过程提供了重要的实验依据。

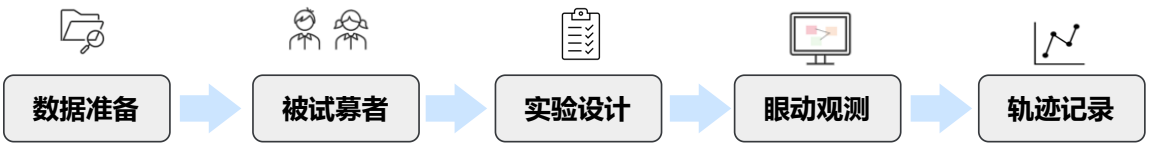


图 3.3　眼动追踪实验

2）标注一致性验证：在实验的最后阶段，我们从五名参与者中随机选取两人对同一子集的数据进行独立标注。随后，针对每个文档文件，我们对比了两名参与者的标注结果。为了确保标注结果的一致性，我们组织五名参与者进行投票，选择最符合多数人预期的标注结果（文档级）。最终选定的标注结果被视为实验的最终数据。

这一流程有效保证了标注结果的客观性和一致性，为后续分析提供了高质量的数据支持。并且所有参与者均具备相关学科背景，并在实验前接受了统一的指导和培训，以确保实验过程的规范性和数据的可靠性。通过上述严谨的实验设计和数据处理流程，我们成功获取了高质量的眼动数据，为深入研究参与者的认知行为提供了有力支持。

（2）研究伦理与参与者权益保护

在研究正式启动前，我们向所有参与者详细说明了实验的目的、流程、潜在风险以及他们的权利，并确保每位参与者充分理解相关内容。我们已按照严格的伦理标准，向相关机构提交了完整的伦理审查申请，并获得了正式的机构伦理批准。这一批准确保了我们的研究流程和方法符合伦理要求，保障了参与者的合法权益。

本研究严格遵循赫尔辛基宣言及其后续修订版本，确保研究的每一个环节都符合国际公认的伦理准则。赫尔辛基宣言是医学研究伦理的基石，旨在保护人类受试者的权益和安全。我们通过严格遵守这些准则，确保研究的科学性与伦理性并重。

为确保参与者的隐私和数据安全，我们在数据处理过程中采取了严格的匿名化措施，移除了所有可能直接或间接识别个人身份的信息。此外，参与者明确同意将其眼动追踪数据用于学术研究目的，并知晓这些数据将仅用于科学分析，不会用于任何商业用途或其他与研究无关的场合。

(3) 人类阅读顺序获取方法

眼动追踪数据处理流程旨在从眼动特征中提取原始人类阅读顺序信息，并将其嵌入到每个边界框中。由于原始数据包含多个特征维度和高采样频率，导致大量重复和缺失数据，并且导出的所有文档的眼动坐标全部在同一个 CSV 文件中，因此需要进行专业化的数据，流程如图 3.4所示，以下是我们数据处理步骤：

1）数据加载与预处理。经数据加载与预处理模块解析原始眼动记录，采用制表符分隔方式读取 CSV 文件，包含注视点索引、空间坐标及事件持续时间等关键字段。通过去除冗余维度、基于 (X,Y) 坐标的去重操作以及首尾无效数据滤除等清洗步骤，确保输入数据的纯净性，并将关键字段转换为整型以提升计算效率。

2）时序分割与存储。通过动态识别注视点索引序列的单调性变化，根据图像个数将连续注视点索引分割到一个独立的文件中。每个文件以标准化命名规则存储为结构化文本文件，保留核心字段，为后续分析提供离散化数据单元。

3）眼动-语义融合：递归遍历 JSON 标注中的边界框语义单元，结合 ±10 像素的

容差缓冲边界，实现眼动坐标与文本区域的匹配。针对部分注视点可能出现偏差或缺失。为了提高数据的准确性，我们采取了以下修正措施：注视点缺失情况一：当注视点缺失时，若其落在已知 OCR 边界框边缘且在一定欧几里得距离范围内，我们使用边缘注视点的序号作为当前边界框的阅读序列索引。注视点缺失情况二：若注视点缺失且不满足情况一的条件，我们使用相邻阅读序列的序号或两个注视点之间的序号进行插值。重复注视点处理：对于多次被眼动返回的注视点，我们仅保留第一次眼动的序号，删除后续重复记录。

4）最终生成增强型 JSON 标注文件，将人类眼动阅读顺序嵌入到视觉富文档数据集的 JSON 文件的每个边界框中。

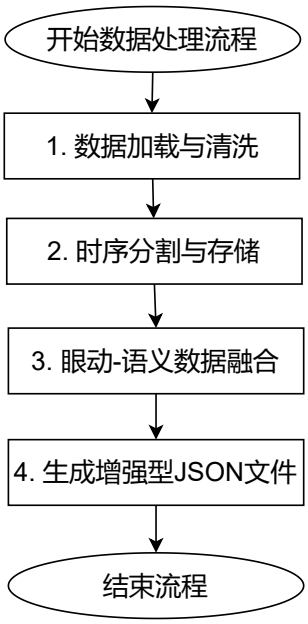


图 3.4　人类阅读顺序获取流程

3.3.2　仿人类的阅读顺序生成

(1) 预排序模型

在预排序模型中，如图 3.5所示，我们通过不同的编码器获取文档不同模态的基本特征，并根据这些特征模拟人阅读顺序。因此，我们提出了四种用于生成阅读顺序的模型。每个模型同时使用单模态或多个模态的信息：**坐标**、**文本**、**文本 + 坐标**和 **文本 + 坐标 + 图像**。

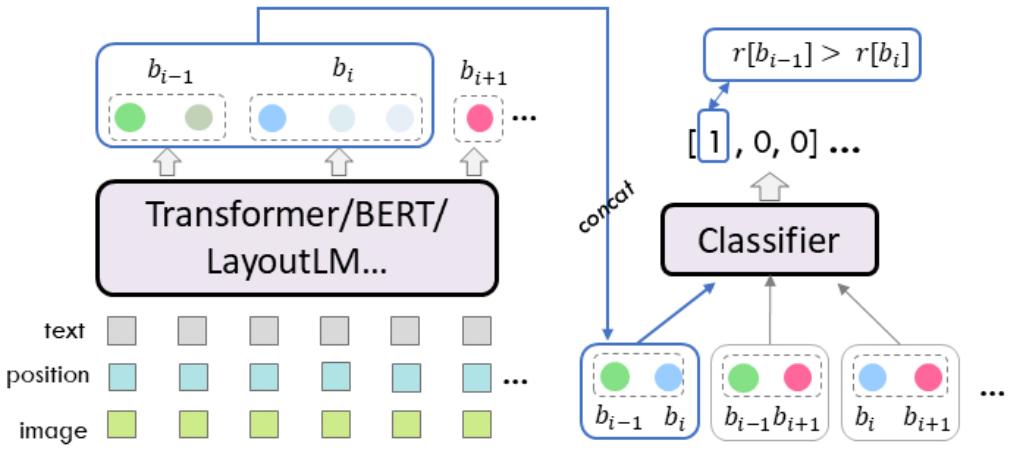


图 3.5 用于相邻边界框的预排序模型，符号">" 表示前者边界框可能在后者边界框之后[78]

每个模型都考虑了影响人类在视觉丰富文档（VRDs）中优先阅读和阅读元素的不同因素，包括元素的位置、元素内的文本以及与之相关的视觉区域。通过使用这些模型，我们可以更准确地评估阅读顺序对人类理解此类文档的影响。

基于坐标的排序。在这个模型中，我们仅使用二维位置来学习每个边界框的顺序。具体而言，首先，我们使用光学字符识别（OCR）工具生成的边界框坐标 $(x_{\text{up}}, y_{\text{up}}, x_{\text{down}}, y_{\text{down}})$（表示左上角和右下角坐标），计算每个边界框 b_i 的中心坐标 (x_i, y_i)。我们将所有边界框的质心坐标两两组合，然后直接将两个组合边界框的中心坐标输入 4 层 Transformer 编码器中，使用了 4 个注意力头，最后输入到全连接层的输入维度为 4 输出维度为 1。这使我们能够预测两个边界框之间的空间关系，即哪一个应该在人类阅读顺序中排在前面或后面。

$$Y = \text{TransformerEncoder}(x_i, y_i, x_j, y_j) \tag{3.3}$$

其中 $Y \in \mathbb{R}^{N \times B \times D}$ 为编码后的序列表示。

$$Z = W \cdot Y + b \tag{3.4}$$

其中 $W \in \mathbb{R}^{D \times 1}$ 为全连接层的权重，b 为偏置项。

基于文本的排序。我们使用 BERT 对两个边框内的文本进行编码（b_i 和 b_j），由于每个边界框包含一个或多个标记，我们取第一个标记位置的隐层表示作为边界框的嵌入。然后我们编码输入到两层全连接层中，最终输出的维度为 1，用于判定文本的前后关系。$X = \{x_1, x2, ..., x_N\} \in \mathbb{N}^{N \times B \times D}$ N 为序列长度，B 为批次大小，D 为

特征维度。公式如下：

$$Y = \text{BERT}(X) \tag{3.5}$$

其中 $Y = \{b_1, b2, ..., b_N\} \in \mathbb{R}^{N \times B \times D}$，为编码后的序列表示，$b_i$ 表示一个文本框中的文本 t_i 的编码结果，然后把编码后的序列特征每两两组合输入到两层神经网络中，公式如下：

$$Z = W^{(2)} \cdot (W^{(1)} \cdot \text{CONCAT}(b_i, b_j) + b^{(1)}) + b^{(2)} \tag{3.6}$$

其中 $W^{(1)} \in \mathbb{R}^{D \times D}$ 为第一层全连接层的权重，$W^{(2)} \in \mathbb{R}^{D \times 1}$ 为第二层全连接层的权重，$b^{(1)}$ 是第一层偏置项，$b^{(2)}$ 为第二层偏置项。

基于文本 + 坐标的排序。为了联合编码文本和二维位置输入，我们使用 LayoutLM，它可以同时对一个文本框中的文本和坐标进行编码表示。与 **文本**部分的操作类似，我们取第一个隐层表示作为分类器的输入。不同之处是这里的 $X = \{x_1, x2, ..., x_N\}$ 中 $x_i = (t_i; x_i, y_i)$, 最终模型可以表示如下：

$$Y = \text{LayoutLM}(X) \tag{3.7}$$

其中 $Y = \{b_1, b2, ..., b_N\} \in \mathbb{N}^{N \times B \times D}$ 为编码后的序列表示，然后将两个边界框编码结果输入给线性层，公式如下：

$$Z = W^{(2)} \cdot (W^{(1)} \cdot \text{CONCAT}(b_i, b_j) + b^{(1)}) + b^{(2)} \tag{3.8}$$

其中 $W^{(1)} \in \mathbb{R}^{D \times D}$ 为第一层全连接层的权重，$W^{(2)} \in \mathbb{R}^{D \times 1}$ 为第二层全连接层的权重，$b^{(1)}$ 是第一层偏置项，$b^{(2)}$ 为第二层偏置项。

基于文本 + 坐标 + 图像的排序。为了实现多模态特征的融合，我们使用 LayoutLMv3 联合编码边界框内的文本、图像和二维位置，一个边界框内包含了三种特征，这里的 $X = \{x_1, x2, ..., x_N\}$ 中 $x_i = (t_i; x_i, y_i, I_i)$，$I_i$ 代表图像特征，具体公式如下所示：

$$Y = \text{LayoutLMv3}(X) \tag{3.9}$$

其中 $Y = \{b_1, b2, ..., b_N\} \in \mathbb{N}^{N \times B \times D}$ 为编码后的序列表示，b_i 表示三种模态经过 LayoutLMv3 编码后的结果，然后将两个边界框的编码结果输入到线性层，公式如下

所示：

$$Z = W^{(2)} \cdot (W^{(1)} \cdot \text{CONCAT}(b_i, b_j) + b^{(1)}) + b^{(2)} \tag{3.10}$$

其中，$W^{(1)} \in \mathbb{R}^{D\times D}$ 为第一层全连接层的权重，$W^{(2)} \in \mathbb{R}^{D\times 1}$ 为第二层全连接层的权重，$b^{(1)}$ 是第一层偏置项，$b^{(2)}$ 为第二层偏置项。我们考虑了文档图像的感兴趣区域（ROIs），分别记为 $\mathbf{I}_i$ 和 $\mathbf{I}_j$。

最后，通过最后激活函数层得到两个边界框内特征位置的前后关系，0 代表 b_i 位于 b_j 后面，1 代表 b_i 位于 b_j 前面。

$$label = \sigma(Z) = \begin{cases} 0, & \text{if } r[b_i] < r[b_j] \\ 1, & \text{if } r[b_i] > r[b_j] \end{cases} \tag{3.11}$$

其中 σ 为 Sigmoid 激活函数，$label \in \mathbb{R}^{T\times B\times 1}$ 为最终输出。b_i 和 b_j 分别表示第 i 个和第 j 个边界框，$r[b_i]$ 表示边界框 b_i 在预测的阅读序列中的前后关系，$r[b_i] < r[b_j]$ 表示 b_i 在阅读顺序序列中出现在 b_j 之前。

（2）损失学习

为了生成适合下游任务的阅读顺序，使用对比损失（Contrastive Loss）来强化模型的学习能力可能是一种更有效的方法。具体而言，我们可以通过多模态编码输出的实体位置关系以及人类的阅读顺序，来生成适合模型的顺序。这一过程同时考虑了人类的阅读顺序以及两个实体之间的“距离”，该距离用符号 $\mathcal{D}$ 表示。具体实现方式如下：给定一对顺序正确的实体和一对顺序错误的实体，模型需要被训练以区分正样本和负样本之间的距离。对比损失函数的定义如下：

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2}\left(y \cdot \mathcal{D}^2 + (1-y) \cdot \max(0, m - \mathcal{D})^2\right) \tag{3.12}$$

其中，y 是标签，用于指示实体之间的相对顺序（1 表示 $\mathbf{e_i}$ 在 $\mathbf{e_j}$ 之前，0 表示 $\mathbf{e_j}$ 在 $\mathbf{e_i}$ 之前）。$\mathcal{D}$ 是两个实体嵌入之间的距离，通常通过欧几里得距离或其他合适的度量方式计算。m 是边际值（margin），用于确保当实体顺序错误时，模型能够学习调整顺序。

具体计算 $\mathcal{D}$ 的方式如下：

$$\mathcal{D} = \|\mathbf{e_i} - \mathbf{e_j}\|_2 = \sqrt{\sum_{k=1}^{n}(e_{i,k} - e_{j,k})^2} \tag{3.13}$$

其中，$\mathbf{e_i}$ 和 $\mathbf{e_j}$ 分别是实体 $\mathbf{e_i}$ 和 $\mathbf{e_j}$ 的嵌入向量，$\|\cdot\|_2$ 表示欧几里得距离（L2 范数）。

通过这种方式，模型能够更好地学习实体之间的相对位置关系，从而学习更符合人类阅读习惯的顺序。这种方法在多模态学习和自然语言处理任务中已被证明是有效的。

（3）基于预排序模型的排序方法

为了对不同模态预排序模型得到的文本框级别的前关系进行排序，我们在基于模型的序列预排序算法中测试了四种原子比较模型。这些模型的主要功能是判断两个元素之间的先后顺序（即“先后顺序”）。预排序算法的核心思想是利用原子比较模型的输出（即 0/1 序列）来构建一个邻接矩阵，从而为后续的排序过程提供基础。具体来说，预排序的实现是基于经典的冒泡排序算法的一种变体。它通过模型判断文本框内容的前后关系，再多次比较和交换序列中相邻的文本框位置，逐步将输入序列重新排列为符合人类阅读习惯的顺序。最终输出经过重新排序的输入序列，使得边界框的排列更加符合人类的自然阅读顺序。具体实现细节和流程，请参见图 3.6。

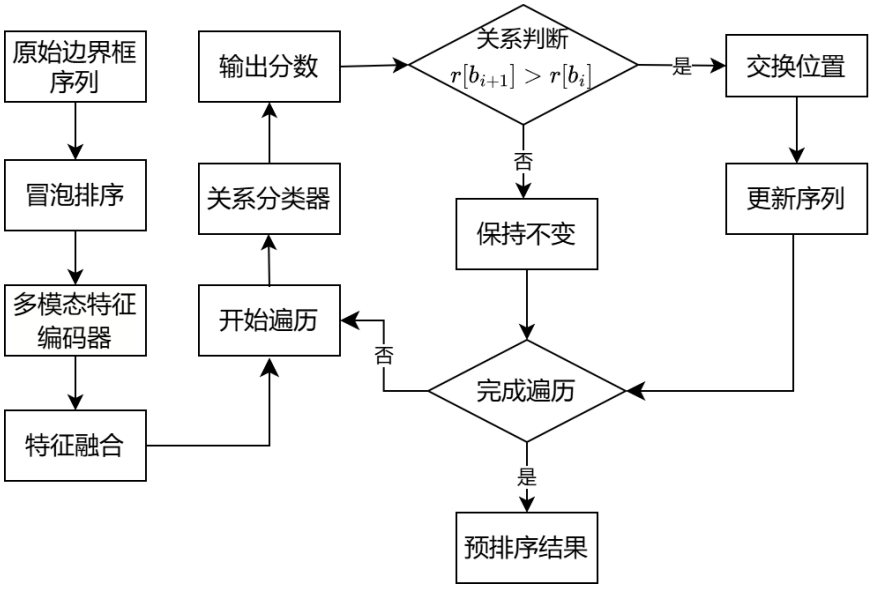


图 3.6　预排序流程图

3.3.3 基于规则的预排序方法

多模态信息提取方法的核心在于对文档序列的精准检测，然而 OCR 引擎输出的不稳定性常导致文本阅读顺序的错乱。受到 gu[95] 和 Li[38] 等人的启发，本文提出的基于坐标进行排序的 Z-order 算法（如算法3.1所示, 后文中提到的 Z-order 都是此算法）遵循人类阅读的基本认知规律，采用基于规则的排序策略：首先依据文本框的垂直位置（y_{tl} ）进行主排序，随后在每行内按照水平位置（x_{tl}）进行次排序。在 XY cut 算法中则采用基于 X 和 Y 轴双坐标的排序机制，结合 Y-X 轴联合搜索策略（含 X 轴搜索与右下搜索）。与其不同的是该算法特别设置了垂直距离阈值机制，当检测到相邻文本框的垂直间距在阈值范围内且后续文本框水平坐标更小时，将自动调整其排列顺序。其中阈值 θ 为垂直邻近阈值（默认 5% 图像高度），通过实验验证对 OCR 错误具有鲁棒性。这种双重排序机制有效克服了 OCR 错误导致的序列错位问题，确保符合 “从左到右、从上到下” 的自然阅读。我们的时间复杂度 O(n²) 但实际收敛快，在 VRD 数据集排序上上比 XYCut 快。

算法 3.1: Z-order **阅读顺序预排序算法**

输入: 原始边界框序列 $B = \{b_1, \ldots, b_n\}$，其中 $b_i = (x_{\text{tl}}^i, y_{\text{tl}}^i, w^i, h^i)$

输出: 符合阅读顺序的重排序序列 B_{sorted}

初步 Y-X 排序，按 y_{tl} 升序排列 B，当 y_{tl} 相等时按 x_{tl} 升序排列

Function ZOrderSort(*B, θ*)**:**

  for $i \leftarrow 1$ **to** $|B| - 1$ **do**

    for $j \leftarrow i$ **to** 0 **do**

      if $x_{tl}^j < x_{tl}^{j-1}$ **且** $|y_{tl}^j - y_{tl}^{j-1}| < \theta$ // 垂直距离小于阈值

      then

        交换 b_j 与 b_{j-1} // 修正行内顺序

        $j \leftarrow j - 1$

      else

        break // 退出当前行调整

  return B_{sorted}

3.3.4 利用仿人类阅读顺序的文档理解模型

（1）基于预训练模型的视觉富文档理解方法

实验使用单模态 BERT 和多模态模型 LAYOUTLMV2 和 LAYOUTLMV3 作为下游任务的主干网络进行，这是一系列基于 Transformer 预训练的视觉富文档理解模型。我们在这个三个模型上进行了微调，具体实验参数可以参见 3.4.4。针对于 SER 任务和 DQA 任务我们的输入都是排序后的序列，其中 BERT 基于 BERT-BASE-UNCASED 的文本信息微调[①]。LAYOUTLMV2/LAYOUTLMV3 输入的是通过将主干替换为 LAYOUTLMV2-BASE 和 LAYOUTLMV3-BASE 进行的微调[②]。

（2）基于大语言模型的视觉富文档理解方法

在 3.3.2 节中，我们介绍了仿人类排序的方法和流程，在 3.4.4 节中我们得到的结论是在四种预排序方法中，基于 LayoutLMv3 的效果更好。因此，在本节中我们使用大语言模型去做 VRUD 任务时只使用了 LayoutLMv3 的排序方法。我们的实验流程有三个阶段：数据预处理，输入顺序的构建和下游任务预测，如图 3.2所示。

首先，我们将通过 OCR 提取文本，坐标和图像三个特征输入给预排序模型，与此同时通过人类眼动追踪技术将图像进行标注，并获取人类阅读顺序。然后我们让排序模型与人类的顺序做对比损失，学习人类的阅读顺序，最终得到了模仿人类的阅读顺序。

最后，我们将排序后的每个实体依次输入到 LLMs 和 VLLMs 中，并通过精心设计的提示（prompt）明确告知这些模型需要执行的具体任务。在 LLMs 的处理过程中，我们仅输入文档中的文本内容 $\mathbf{t}_i$ 以及对应的位置坐标 $\mathbf{x}_i$ 和 $\mathbf{y}_i$，这些坐标信息用于提供文本在文档中的空间位置上下文，从而帮助模型更好地理解文本的结构化信息。而在 VLLMs 的处理过程中，除了输入文本内容 $\mathbf{t}_i$ 和位置坐标 $\mathbf{x}_i$ 和 $\mathbf{y}_i$ 外，我们还会额外输入视觉特征 $\mathbf{I}_i$，这些视觉特征通常是从文档图像中提取的，能够为模型提供更丰富的视觉上下文信息，例如字体顺序、布局结构等。通过结合文本、位置和视觉特征，VLLMs 能够更全面地理解文档内容，从而在任务执行中表现出更强的多模态推理能力。这种分阶段的输入设计不仅充分利用了 LLMs 在文本理解上的优势，还通过 VLLMs 引入了视觉信息的补充，使得模型在处理复杂文档任务时能够更

① https://github.com/google-research/bert
② https://github.com/microsoft/unilm

加精准和高效。具体公式如下：

$$\mathbf{L}_i = \text{LLMs}(\mathbf{t}_i; \mathbf{x}_i, \mathbf{y}_i; \text{Prompt}_1) \tag{3.14}$$

在 VLLMs 中加入了视觉信息 I_i，具体公式如下：

$$\mathbf{L}_i = \text{VLLMs}(\mathbf{t}_i; \mathbf{x}_i, \mathbf{y}_i; \mathbf{I}_i; \text{Prompt}_2) \tag{3.15}$$

SER 提示：在语义实体识别（Semantic Entity Recognition, SER）任务中，模型的目标是根据 OCR 引擎输出的文本段内容和空间坐标对其进行分类，而无需额外的注释。为了确保模型能够正确理解任务，我们提供了基本的指示，明确说明目标是为每个文本段分配一个类别标签。此外，我们还包含了一个简单且不相关的示例，以帮助模型理解输入和输出的格式。

例如，给定文本段 “Port Of Loading:”，模型应将其分类为 “question”，而 “San Francisco, California” 则应分类为 “answer”。通过这种方式，模型能够生成一致的输出格式，如下所示：{“Port Of Loading:”:“question”, “San Francisco, California”: “answer” } 这种格式化的输入和输出不仅帮助模型理解任务要求，还确保了分类结果的一致性。通过这种方式，模型能够有效地处理复杂的文档内容，并生成结构化的输出，从而提升语义实体识别的准确性。

DQA 提示：在文档视觉问答（Documents Question Answering, DQA）任务中，模型接收用户查询并基于结构化的提示生成响应。这种提示机制确保了模型的回答与用户查询高度相关，从而提高了答案的准确性。

例如，当用户查询 “What is the official language of Spain?” 时，模型将生成如下响应：{“answer”: “Spanish”} 如果模型在文档中未找到相关信息，则响应为：{“answer”: “No”}。这种方法通过将文档内容与用户查询紧密结合，显著提升了模型在视觉富文档（VRDs）理解任务中的性能。值得注意的是，用于格式化的示例在测试数据集中从未出现过，从而避免了模型对特定示例的过拟合，确保了模型的泛化能力。

通过这种结构化的提示机制，模型能够更准确地理解用户查询，并从文档中提取相关信息，从而生成高质量的答案。这种方法不仅提高了模型的性能，还增强了其在复杂文档理解任务中的实用性。

3.3.5 人类阅读顺序相似性评估方法

在序列相似性分析中，斯皮尔曼相关系数（Spearman's Rank Correlation Coefficient）和肯德尔相关系数（Kendall's Tau）是两种常用的非参数统计方法，主要用于衡量两个序列之间的单调关系[96-97]。这两种方法特别适合处理非线性关系和秩次数据，能够有效评估序列之间的相似性，且无需假设数据服从正态分布。主要应用场景有以下三种：**1）评估序列的一致性**：通过计算两个序列之间的斯皮尔曼或肯德尔相关系数，可以量化它们的相似性。例如，在时间序列分析中，可以用来评估两个时间序列之间的趋势一致性；**2）处理非线性关系**：这两种系数对非线性关系的敏感性较高，适用于分析非线性相关的序列数据；**3）处理秩次数据**：当数据为秩次形式（如排名数据）时，斯皮尔曼和肯德尔相关系数比传统的皮尔逊相关系数更为适用；一般来说肯德尔系数适用于衡量两个序列的排序一致性，尤其在数据中存在较多重复值或需要衡量局部排序一致性时表现优异；斯皮尔曼系数适用于衡量两个序列的单调关系，能够捕捉序列之间的全局趋势。

（1）肯德尔系数（Kendall's Tau）肯德尔系数是一种用于衡量两个序列之间排序一致性的非参数统计量。其定义如下：给定两个序列 $X = (x_1, x_2, \ldots, x_n)$ 和 $Y = (y_1, y_2, \ldots, y_n)$，肯德尔系数 τ 的计算公式为：

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(x_i - x_j) \cdot \text{sgn}(y_i - y_j), \tag{3.16}$$

其中：

- $\text{sgn}(\cdot)$ 是符号函数，当 $x_i > x_j$ 时取值为 $+1$，当 $x_i < x_j$ 时取值为 -1，当 $x_i = x_j$ 时取值为 0；
- n 是序列的长度；
- 肯德尔系数的取值范围为 $[-1, 1]$：$\tau = 1$ 表示两个序列完全一致；$\tau = -1$ 表示两个序列完全相反；$\tau = 0$ 表示两个序列无相关性。

（2）斯皮尔曼系数（Spearman's Rank Correlation Coefficient）斯皮尔曼系数是一种用于衡量两个序列之间单调关系的非参数统计量。其定义如下：给定两个序列 X 和 Y，首先将它们的值转换为秩次（rank），记为 R_X 和 R_Y。斯皮尔曼系数 ρ 的计算公式为：

$$\rho = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3.17}$$

其中：

- $d_i = R_X(i) - R_Y(i)$ 是第 i 个元素的秩次差；
- n 是序列的长度；
- 斯皮尔曼系数的取值范围为 $[-1, 1]$：$\rho = 1$ 表示两个序列完全单调正相关；$\rho = -1$ 表示两个序列完全单调负相关；$\rho = 0$ 表示两个序列无单调关系。

3.4 实验与分析

3.4.1 数据集

实验数据集：我们在实验中使用了两个公开的开源数据集 FUNSD[44]和经过挑选的 InfographicVQA[98]，以及一个内部收集的数据集 SEAB。这些数据集的详细统计信息如表 3.1所示。

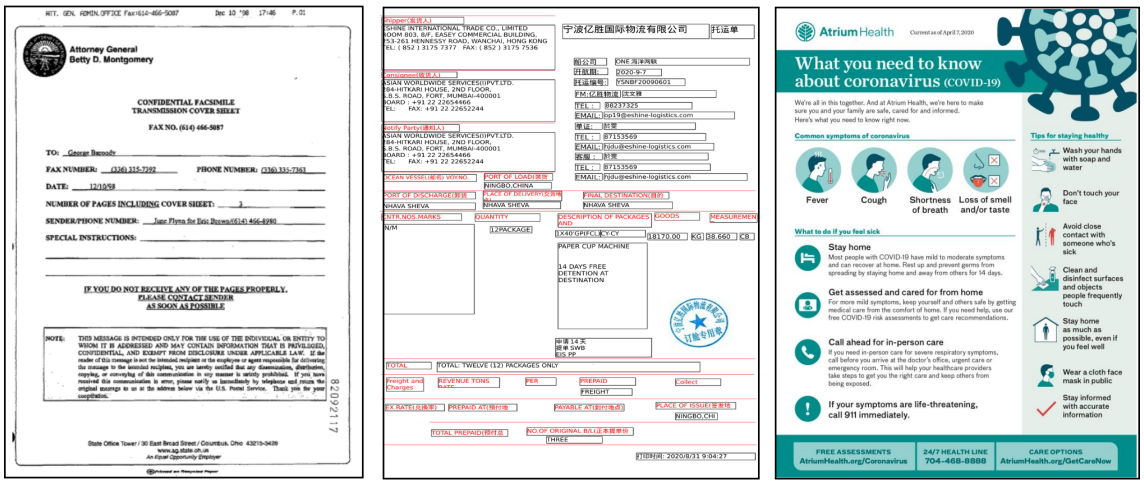
表 3.1 使用的数据集统计信息，包括文档数（Doc）、边界框数（BD）和标记数（Token）

数据集	# 训练			# 测试		
	文档数	边界框数	标记数	文档数	边界框数	标记数
FUNSD	149	7,441	22,512	50	2,332	8,973
SEAB	160	10,024	16,055	50	3,430	7,022
InfoGraph	100	12,650	24,364	30	3,794	7,308
TOTAL	409	30,115	62,931	130	9,556	23,123

- **FUNSD**：此数据集由 199 个经过精细标注的扫描表单组成。每个语义实体都分配了一个唯一的标识符（id）、一个标签（如 “Question”、“Answer”、“Header” 或 “Other”）、一个边界框、一个指向其他实体的链接列表以及一个单词列表。具体示例参见图 3.7（a）。
- **SEAB**：此数据集从国际海运场景中提取，包含 160 个训练文档和 50 个测试文档。数据由 PDF 图像组成，这些图像基于 PPOCRLabel[①]标注了文本、位置和类型信息。数据具有粗粒度的三个标签（“Question”、“Answer” 或 “Other”）和细粒度的 56 个标签（如发货人、发货人-value、起运港、起运港-value 等）具体示例参见图 3.7（b）。

① https://github.com/Evezerest/PPOCRLabel

- **InfoGraph**：InfoGraph 是一个专门为视觉问答任务设计的数据集，旨在通过自动理解信息图表（InfographicVQA）中的多模态信息（如文本、图形、布局等）来回答问题。信息图表是一种通过结合文本、图形和视觉元素来传达信息的文档形式，其复杂的布局和多模态特性使得自动理解成为一项具有挑战性的任务。它包含了多样化的信息图表集合以及对应的问题-答案。这些信息图表中的问题需要模型能够联合推理文档布局、文本内容、图形元素和数据可视化信息。我们从该数据集中选取了包括 100 张用于训练集和 30 张用于测试集。具体示例参见图 3.7（c）。



(a) (b) (c)

图 3.7 FUNSD（a）、SEAB（b）和 InfoGraph（c）数据集的样例展示

3.4.2 评价指标

在 SER 任务中，混淆矩阵（Confusion Matrix）如表 3.2 是评估模型性能的重要工具。通过混淆矩阵，我们可以计算出 Precision（精确率）、Recall（召回率）和 F1 分数，这些指标能够帮助我们更全面地评估分类任务的性能。

精确率（Precision）是预测为正例中实际为正例的比例，其公式为：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.18}$$

召回率（Recall）是实际为正例中被正确预测为正例的比例，其公式为：

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.19}$$

F1 分数（F1 Score）是精确率和召回率的调和平均数，公式为：

$$F1 = \quad = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \times \text{Recall}} \tag{3.20}$$

表 3.2　混淆矩阵

真实情况	**预测结果**	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

3.4.3　顺序相似性评估结果

图 3.8 展示了四种不同的模型在三个数据集上的序列相关性评估结果。研究结果揭示了以下几点：首先，结合文本、边界框和图像信息的模型表现最佳，其生成的阅读顺序与人类阅读顺序的相关性最高。这表明，多模态特征（包括文本、边界框和图像）的联合使用可以显著提高模型对文档阅读顺序的建模能力。其次，仅依赖边界框信息的模型表现最差，其生成的阅读顺序与人类阅读顺序的相关性最低。这表明，仅依赖边界框信息无法充分捕捉文档的复杂结构，尤其是在视觉富文档（VRD）中。第三，结合文本和边界框信息的模型以及仅使用文本信息的模型表现介于前两者之间。这表明文本信息在阅读顺序生成中起到了重要作用，但仅依赖文本信息仍不足以完全模拟人类的阅读行为。

此外，研究还观察到人类在阅读不同类型文档时的行为差异。在阅读弱表格结构的视觉富文档时，人类不仅依赖文本和位置信息，还会利用字体、背景等视觉特征辅助理解文档的结构和内容，显示视觉特征在这类文档中起到了重要的辅助作用。在阅读具有明显表格结构的文档时，人类主要依赖文本和位置信息来理解文档内容，视觉特征的作用相对较弱，说明在表格结构清晰的情况下，基本的文本和位置信息足以支持有效的阅读顺序生成。信息图表通常包含丰富的图像和视觉元素，因此需要模型能够同时捕捉文本、位置和视觉特征。实验结果表明，仅依赖文本或边界框信息的模型在处理信息图表时表现较差，而结合多模态特征的模型能够更好地模拟人类的阅读行为。

实验结果验证了多模态特征在文档阅读顺序生成中的重要性。具体来说：文本信息提供了文档的主要内容，是阅读顺序生成的基础。边界框信息反映了文档的布

局结构，有助于模型理解文本和图像的位置关系。图像信息在视觉富文档和信息图表中，能够提供额外的视觉线索，帮助模型更好地理解文档的语义和结构。总结来说，通过对比机器生成的输入序列与人类的阅读序列，我们发现，多模态特征（文本、边界框和图像）的联合使用能够显著提升模型在阅读顺序生成任务中的性能，尤其是在处理视觉富文档和信息图表时，多模态特征的引入能够更好地模拟人类的阅读行为。未来的研究可以进一步探索如何更有效地融合多模态特征，以提升模型在复杂文档理解任务中的表现。

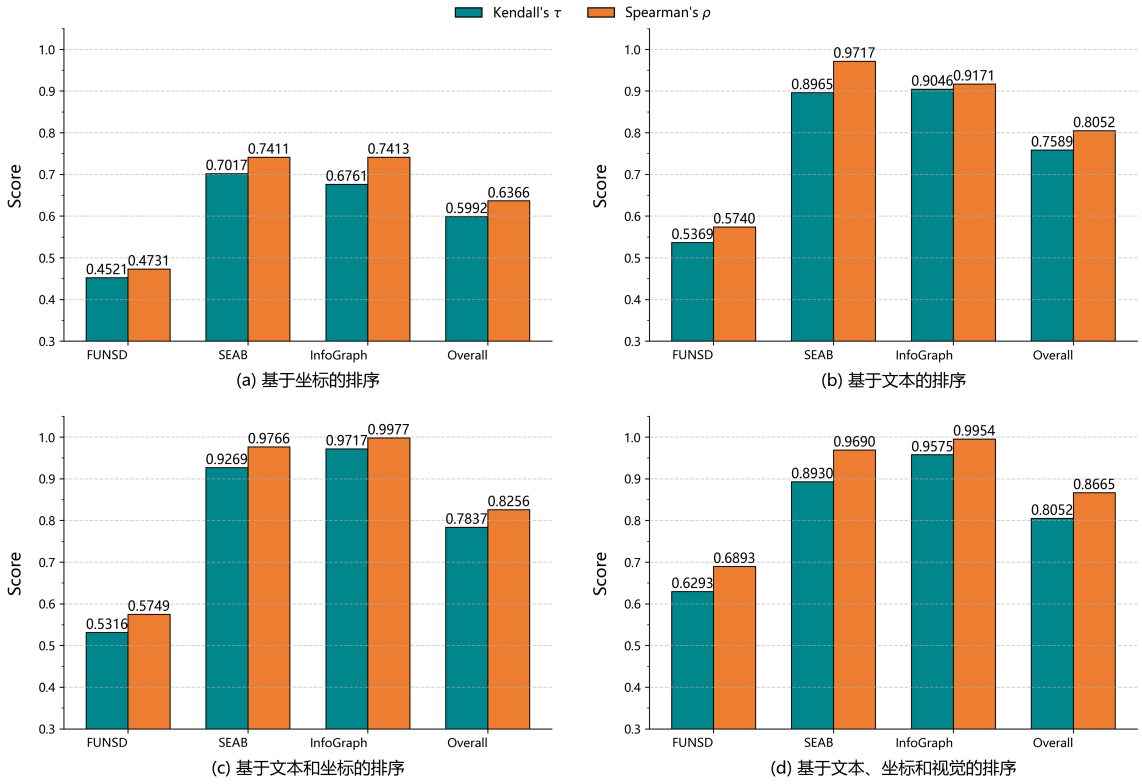


图 3.8　人类顺序的相似性与仿人类顺序相似性比较

3.4.4　预训练模型的实验结果

1. 硬件环境:

本章实验所使用的硬件环境包括两张 Geforce RTX 3090 Ti 显卡，系统内存容量为 128GB。软件环境配置如下：操作系统为 Ubuntu 20.04.2，深度学习框架为 Pytorch 1.8.1，transformer 库版本为 4.5.0，CUDA 版本为 11.4，编程语言为 Python 3.8.10，如表 4.1所示。

2. 参数设置

表 3.3　实验的软硬件环境

部件	**参数**
操作系统	Ubuntu 20.04.2
系统内存	128G
CPU 处理器	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz
GPU 处理器	Geforce RTX 3090 Ti
Python 版本	1.8.1
Pytorch 版本	3.8.10
CUDA 版本	11.4
transformer 版本	4.5.0

所有模型使用 Adam 优化器进行训练，权重衰减率设置为 1×10^{-2}，批量大小设置为 8。对于 FUNSD 数据集，我们将 Dropout 设置为 0.3，学习率为 5×10^{-5}，以防止模型过拟合，训练 20 次迭代以达到收敛。而对于 SEABILL 数据集，我们将 Dropout 降低到 0.1，学习率为 5×10^{-5}，我们将模型训练 50 次迭代以达到收敛，并实现更稳定的性能。在 DQA 任务中，InfoGraph 数据集训练批量大小设置为 8，为了更好的拟合任务，学习率为 1×10^{-5}，训练 20 次迭代。

3. 主要实验结果

本研究的目的是利用语义实体识别（Semantic Entity Recognition，简称 SER）任务分析阅读顺序对视觉富文档理解（Visually Rich Document，简称 VRD）的影响，研究对象为文档的 FUNSD 和 SEAB 子集。表 3.4 展示了结果。我们发现，阅读顺序对 SER 任务的影响取决于所使用的文档 AI 模型。当使用 BERT 模型时，简单的 Z 顺序效果最好，每种顺序的效果都优于原始顺序。在使用多模态的 LayoutLMv2 和 LayoutLMv3 模型时，多模态排序效果最佳，大多数结果略优于 Z-order。这些结果表明，人类阅读顺序和机器排序对 SER 任务的效果具有显著的提升，且不同模型对这些顺序的敏感程度不同。

文档问答（Document Question Answering，简称 DQA）是视觉富文档理解（Visually Rich Document，简称 VRD）中的一项挑战性任务，它要求机器理解图像中的视觉和文本内容，并回答有关问题。表 3.4 列出了仅使用文本的基线 BERT、布局感知的多模态基线 LayoutLMv2 和 LayoutLMv3 在 InforGraph 子集上的平均归一化 Levenshtein 相似度（Average Normalized Levenshtein Similarity，简称 ANLS）分数。我们观察到 LayoutLMv2 和 LayoutLMv3 模型的表现大幅超越了仅使用文本的基线（BERT）。虽

然整合人类阅读顺序可以增强现有文档 AI 模型在下游任务中的表现，但这并不总是优于使用基于规则的方法生成的类似人类阅读顺序。这表明，真正的人类阅读顺序可能并非提升现有机器文档 AI 模型所必需的。出现这种情况有几个可能的原因。首先，现有多模态文档模型所使用的大多数数据集都是按照简单规则排序的，因此更适合使用诸如 Z 模式等简单规则生成的顺序。此外，个别人类的阅读顺序可能非常嘈杂，除非通过收集大量的人类眼动数据构建大型人类眼动数据集。

从整体结果来看，在四种多模态仿人类顺序生成的模型中，文本 + 坐标 + 图像特征排序在下游任务上的结果是最好的，这个结论也证明传统的预训练模型在理解时视觉富文档需要多种模态的协同。

表 3.4 在 DocTrack 数据集上的结果。OCR 代表原始顺序，Human 是人类的眼动阅读顺序，Z-order 和 XYcut[38] 是由专家经验生成的两种顺序，Model-B、Model-T、Model-T+B 和 Model-T+B+I 分别代表基于 Box、Text、Text+Box 和 Text+Box+Image 等原子比较模型。R/H/M 分别指的是由规则、人类和模型生成的顺序

	Preorder	Modality			Type	FUNSD			SEAB			InfoGraph
		Text	Pos	Img	R/H/M	P (%) ↑	R (%) ↑	F1 (%) ↑	P (%) ↑	R (%) ↑	F1 (%) ↑	ANLS (%) ↑
BERT	Human	✓	✓		H	60.52	60.77	60.47	60.75	60.83	60.79	4.65
	OCR	✓			R	56.69	62.11	60.33	58.99	60.01	59.51	3.82
	Z-order	✓	✓		R	**64.09**	**65.28**	**64.66**	**63.44**	**62.78**	**63.11**	**5.88**
	XYcut	✓	✓		R	60.16	60.84	60.19	59.24	60.08	59.65	3.71
	Model_B	✓	✓		M	60.19	61.98	60.92	59.01	60.98	59.98	2.94
	Model_T	✓			M	62.80	63.16	62.87	61.04	60.22	60.62	2.99
	Model_{T+B}	✓	✓		M	61.30	62.43	61.80	60.52	60.77	60.47	3.14
	Model_{T+B+I}	✓	✓	✓	M	62.74	64.51	63.45	62.43	62.17	62.80	3.21
LayoutLMv2	Human	✓	✓		H	87.38	83.82	85.41	77.44	74.39	75.88	15.69
	OCR	✓			R	86.94	80.95	83.44	78.56	73.02	75.69	12.50
	Z-order	✓	✓		R	**88.00**	84.46	86.06	78.05	74.77	76.37	**18.09**
	XYcut	✓	✓		R	84.01	83.12	83.55	75.01	**78.41**	76.61	12.38
	Model_B	✓	✓		M	87.23	85.16	86.13	78.45	73.01	75.63	12.98
	Model_T	✓			M	85.35	82.40	83.77	77.64	77.14	77.39	12.67
	Model_{T+B}	✓	✓		M	86.76	86.61	86.57	**80.24**	74.84	77.45	14.70
	Model_{T+B+I}	✓	✓	✓	M	87.00	**87.01**	**86.98**	79.74	75.58	**77.60**	14.70
LayoutLMv3	Human	✓	✓		H	91.47	91.19	91.33	69.22	65.57	67.35	18.64
	OCR	✓			R	90.96	92.00	91.48	72.96	66.71	69.70	18.22
	Z-order	✓	✓		R	**94.63**	92.85	**93.73**	**77.27**	68.24	72.47	**20.58**
	XYcut	✓	✓		R	90.10	89.69	89.90	71.05	66.73	68.82	17.63
	Model_B	✓	✓		M	92.37	91.95	92.16	75.81	70.00	72.79	17.76
	Model_T	✓			M	91.86	93.09	92.47	74.01	67.56	70.64	17.52
	Model_{T+B}	✓	✓		M	92.39	92.85	92.62	77.17	**70.98**	**73.94**	18.01
	Model_{T+B+I}	✓	✓	✓	M	93.98	**93.14**	93.56	74.17	68.97	71.47	18.24

3.4.5 大语言模型的实验结果

在本小节及以后的章节中我们将多模态排序模型统一使用 Mimic Human 代替。

SER 任务：命名实体识别（Semantic Entity Recognition, SER）是信息抽取领域的一个重要子任务，其目标是从文档中定位命名实体并将其分类到预定义的类别中。在本研究中，我们在 FUNSD 和 SEAB 数据集上进行了实验，这两个数据集包含四种类别：Header（标题）、Question（问题）、Answer（答案）和 Other（其他）。我们在 GPT-3.5 和 GPT-4.0 模型上进行了测试，实验结果如表 3.5所示。实验结果表明，顺序在两个模型的六个 F1 分数结果中，Z-order 有五个取得了最高的 F1 值，而次优结果则是由多模态模型模仿人类阅读顺序生成的顺序所取得。这点符合人类阅读顺序的特点，因为人类阅读在特定的文档结构中也是遵循 Z-order 顺序。特别值得注意的是，通过模型生成的顺序在多个结果中均优于其他顺序，尤其是在 FUNSD 数据集上，其平均 F1 分数比默认顺序高出十几个百分点。这一发现进一步验证了引入人类阅读顺序对提升模型性能的重要性，证明了仿人类阅读顺序在结合人类阅读顺序的同时结合了文档本身特点时对大模型理解视觉富文档理解效果有显著提升。总的来说，Mimic human 顺序在命名实体识别任务中表现出色，尤其是在处理复杂文档时，能够显著提升模型的准确性和鲁棒性。

表 3.5　不同阅读顺序在 SER 任务上的结果，其中 Mimic Human 是模仿人类的阅读顺序

Model	ORDER	Type	FUNSD			SEAB		
		R/H/M	P (%)↑	R (%)↑	F1 (%)↑	P (%)↑	R (%)↑	F1 (%)↑
LLaMA3 8B	OCR	R	53.00	29.44	37.85	36.43	77.86	49.64
	Human	H	74.63	36.02	48.59	**40.30**	83.30	**54.32**
	Mimic Human	M	79.72	44.53	57.14	39.09	**87.10**	53.96
	Z-order	R	**81.61**	**59.43**	**68.77**	37.57	79.89	51.46
GPT-3.5	OCR	R	69.93	41.83	52.34	31.21	58.04	40.59
	Human	H	78.28	53.44	63.52	31.04	56.17	39.98
	Mimic Human	M	76.58	62.33	68.72	31.28	59.56	41.02
	Z-order	R	**78.88**	**66.59**	**72.21**	**31.82**	**63.01**	**41.98**
GPT-4.0	OCR	R	83.45	61.67	70.93	39.45	45.38	42.76
	Human	H	85.86	63.63	73.09	47.49	59.96	53.00
	Mimic Human	H	**87.74**	74.06	80.64	46.50	59.06	52.03
	Z-order	R	87.11	**75.31**	**80.78**	**50.47**	**67.62**	**57.81**

DQA 任务：文档问答（Document Question Answering, DQA）任务要求模型根据

输入的文档信息和问题生成相应的答案，是视觉富文档理解中的一项重要任务。在本次实验中，我们基于 InfographicVQA 数据集的实验数据，并通过引入不同的阅读顺序（如人类阅读顺序和模型生成的顺序）来验证顺序对 DQA 任务的影响。我们使用 ANLS（Average Normalized Levenshtein Similarity）和 F1 分数作为评估指标来衡量模型的表现。实验结果如表 3.6所示，我们在 GPT-3.5、GPT-4.0、GLM-4V 和 Qwen2-VL-7B 四个模型上进行了实验。实验结果表明，Z-order 顺序在所有模型中 ANLS 分数均取得了最佳效果，其次是仿人类阅读的顺序。与此同时 Z-order 和仿人类阅读顺序的结果要优于人类阅读顺序以及远高于原始 OCR 的顺序，这一结果进一步证明了阅读顺序对 DQA 任务具有显著影响。并且与我们在命名实体识别（SER）任务中的结论一致，支持了我们的核心观点：通过引入人类阅读顺序或优化的模型生成顺序，可以显著提升视觉富文档理解任务的性能。

表 3.6　不同阅读顺序在文档问答（DQA）任务上的结果

Model	**Order**	**Type(R/H/M)**	**Modality**	**F1 (%)↑**	**ANLS (%)↑**
LLaMA3 8B	OCR	R	T+B	18.51	18.51
	Human	H	T+B	19.49	26.07
	Mimic Human	M	T+B	20.78	26.78
	Z-order	R	T+B	**22.65**	**31.67**
GPT-3.5	OCR	R	T+B	22.01	26.45
	Human	H	T+B	25.49	31.51
	Mimic Human	M	T+B	24.68	31.57
	Z-order	R	T+B	**28.43**	**33.33**
GPT-4.0	OCR	R	T+B	31.08	38.76
	Human	H	T+B	36.22	43.55
	Mimic Human	M	T+B	35.71	44.19
	Z-order	R	T+B	**37.49**	**44.53**
GLM-4V 9B	OCR	R	T+B+I	43.57	62.06
	Human	H	T+B+I	42.82	64.81
	Mimic Human	M	T+B+I	**43.71**	64.91
	Z-order	R	T+B+I	43.10	**65.04**
Qwen2-VL 7B	OCR	R	T+B+I	41.38	63.81
	Human	H	T+B+I	40.65	64.29
	Mimic Human	M	T+B+I	41.78	65.36
	Z-order	R	T+B+I	**42.50**	**66.29**

图 3.9　上图展示了模拟人类阅读顺序，下图展示了默认的 OCR 顺序

输入顺序对大型语言模型（LLMs）预测结果的影响的可视化分析：为了更清晰的分析阅读顺序对视觉富文档理解任务的影响，我们可视化了大模型预测输入不同顺序实体时的标签概率。在图 3.9中，我们展示了输入默认 OCR 顺序和 mimic human 顺序的结果，由于 header 类型的实体一般位于文档的顶部位置，如果是 OCR 默认顺序，输入可能靠后，大模型会错误的把 header 类型预测为别的实体类型。如 3.9右图中的 “ATT. GEN. ADMIN OFFICE” 和 “Dec 10 ’98 17 :46P 01” 被错误的预测为”question” 和 “answer” 类型。

3.5　本章小结

本章首先介绍了通过眼动追踪记录仪记录人类眼动信息，并通过后续的处理完成了基准数据集的构建，在此基础上提出了分别基于坐标、文本、文本 + 坐标和文本 + 坐标 + 图像四种多模态预排序模型，并将人类阅读顺序融入到多模态预排序模型中，将模仿的人类阅读顺序结果和人类阅读顺序通过肯德尔相关系数和斯皮尔曼

相关系数进行了相似性评估。最后我们通过传统的多模态预训练模型和提示大语言模型两种方法进行了验证，在 SER 任务和 DQA 任务上的结果都表明我们提出的预排序模型的有效性。

第四章　基于眼动追踪与热图提示的仿人类注意力机制学习方法

4.1　引言

文档智能技术的飞速发展极大地改变了我们获取、处理和理解信息的方式。富视觉文档作为一种包含表格、图形、图表和其他视觉元素的信息载体，因其能够以紧凑的形式传递大量信息而得到广泛应用。然而，理解这些文档并不仅仅依赖于文本数据的处理，还需要整合视觉元素、布局结构，并结合与人类感知和认知过程相匹配的机制。

富视觉文档（VRD）通过结合文本、图形和视觉元素，能够以直观的方式传递复杂的信息。然而，这种多模态特性也使得自动理解 VRD 成为一项具有挑战性的任务。与传统的纯文本文档不同，VRD 的理解需要模型能够同时处理文本、图像、布局等多种信息，并理解它们之间的语义关系。

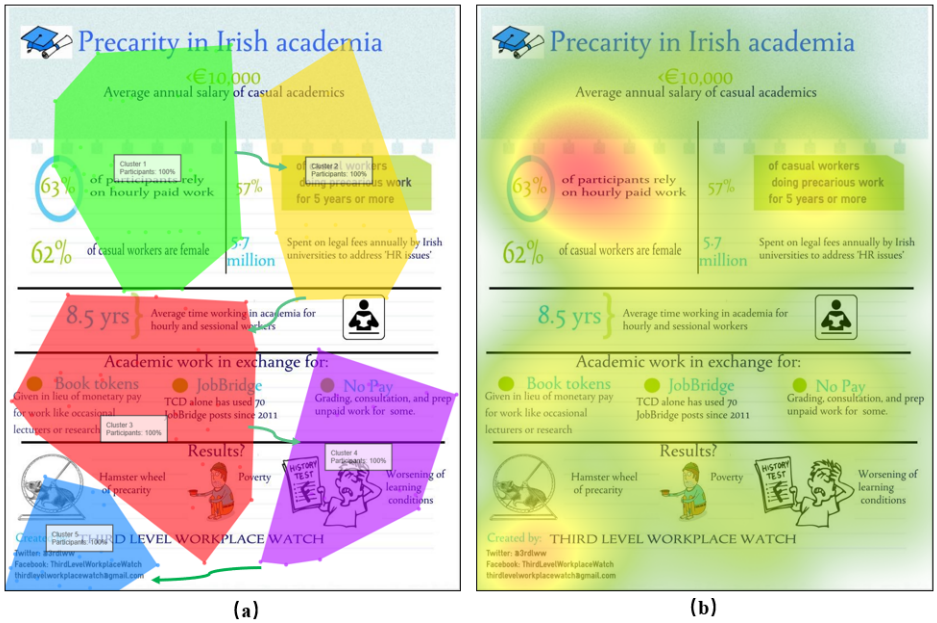


图 4.1　人类阅读视觉关系图（VRD）的行为可视化：(a) 阅读顺序和 (b) 视觉热图

尽管大型语言模型（LLMs）和视觉大语言模型（VLLMs）[59,63,99]在文本理解和图像处理方面取得了显著进展，但它们在处理 VRD 时仍难以复现人类的深度认知能力。现有的模型虽然在文本生成、问答系统等任务中表现出色，但在处理复杂的多

模态文档时，往往无法像人类一样高效地整合文本、视觉和布局信息。例如，虽然一些研究（如 LayoutReader[37] 和 DocTrack[78]）尝试通过模拟人类的阅读顺序来提升模型性能，但这些方法仅捕捉到了人类认知行为的一小部分。人类的阅读行为并不仅仅是线性的，而是受到视觉注意力、上下文理解等多种认知机制的共同驱动。此外，现有的 AI 模型缺乏类似人类的视觉注意力机制，无法在复杂文档中快速聚焦于关键信息区域[100-101]，导致其在处理视觉富文档时表现不佳。

人类的认知机制在解读 VRD 方面发挥着至关重要的作用。如图 4.1 所示，人类读者通过阅读顺序（Reading Order）[102-103]和视觉注意力（Visual Attention）[104-105]等认知策略，能够高效地浏览复杂的文档布局。这些机制是人类长期进化的结果，使得我们能够在阅读过程中快速确定信息的优先级，并整合文本、视觉和布局数据。

阅读顺序：人类在阅读文档时，通常会遵循一定的逻辑顺序，例如从左到右、从上到下[106]，或者根据文档的布局结构（如标题、段落、图表等）进行信息获取。

视觉注意力：人类能够通过视觉注意力机制，快速聚焦于文档中的关键区域，忽略不相关的信息。这种选择性注意力机制使得我们能够在复杂文档中高效地提取信息。

为了提升 AI 模型对 VRD 的理解能力，模仿人类的认知行为可能是一个重要的研究方向[107]。具体而言，可以通过引入人类的阅读顺序策略，使模型更好地理解文档的逻辑结构，从而提高信息提取的准确性。同时，模拟人类的视觉注意力机制，使模型能够快速聚焦于文档中的关键区域，优先处理重要信息。此外，结合文本、图像和布局信息的多模态融合技术，可以进一步提升模型对复杂文档的理解能力。

我们总结了以下三个贡献：我们提出了一种将人类视觉注意力注入到大型语言模型和视觉语言模型中的方法，使得这些模型能够更精准地复现人类的阅读行为。我们证明了更精细的 Z-order 阅读顺序（大型语言模型和视觉语言模型更青睐这种顺序）比传统的人类阅读顺序更适配机器模型。实验表明，使人工智能模型与人类的认知模式（如阅读顺序和注意力区域）对齐，可以显著提升大型语言模型和视觉语言模型在版面理解任务上的性能。这凸显了认知启发式注意力机制在提升人工智能版面理解能力方面的潜力，强调了使机器模型与人类认知过程相协调对更有效地理解文档的重要性。

4.2 框架对比

当前，视觉富文档信息抽取领域取得了重要突破，其中多模态文档理解模型发挥了关键作用[5-10,40]。这类模型通常采用”预训练-微调”的范式来实现特定领域的任务处理。与之形成对比的是，大语言模型（LLMs）采用了”预训练-提示”的创新模式，通过精心设计的提示语即可引导模型产生目标输出，显著降低了对标注数据的依赖。图 4.2直观呈现了这两种方法在文档理解任务中的不同特点。

传统文档理解模型，如 LayoutLMv2 和 LayoutLMv3，必须经过完整的训练数据集微调后才能进行推理应用。而 LLMs 则突破了这一限制，跳过了复杂的微调环节，直接以模板化的提示作为输入（如 GPT4[53]、Qwen2-VL[63]等），进而获取预期结果。在这一机制中，提示语的设计尤为关键，其核心要素包括指令部分和查询部分，指令部分作为模型的前导信息，包含任务说明、标签对应关系、限制条件等要素，并可选择性添加示例样本和输出格式示范；查询部分是为了聚焦具体问题，由问题主体和详细问题描述构成。

通过这种结构化的提示方式，LLMs 能够生成规范化的输出结果，为后续任务处理提供可靠的基础。这种创新方法不仅简化了模型应用流程，还显著提升了模型的适应性和灵活性。

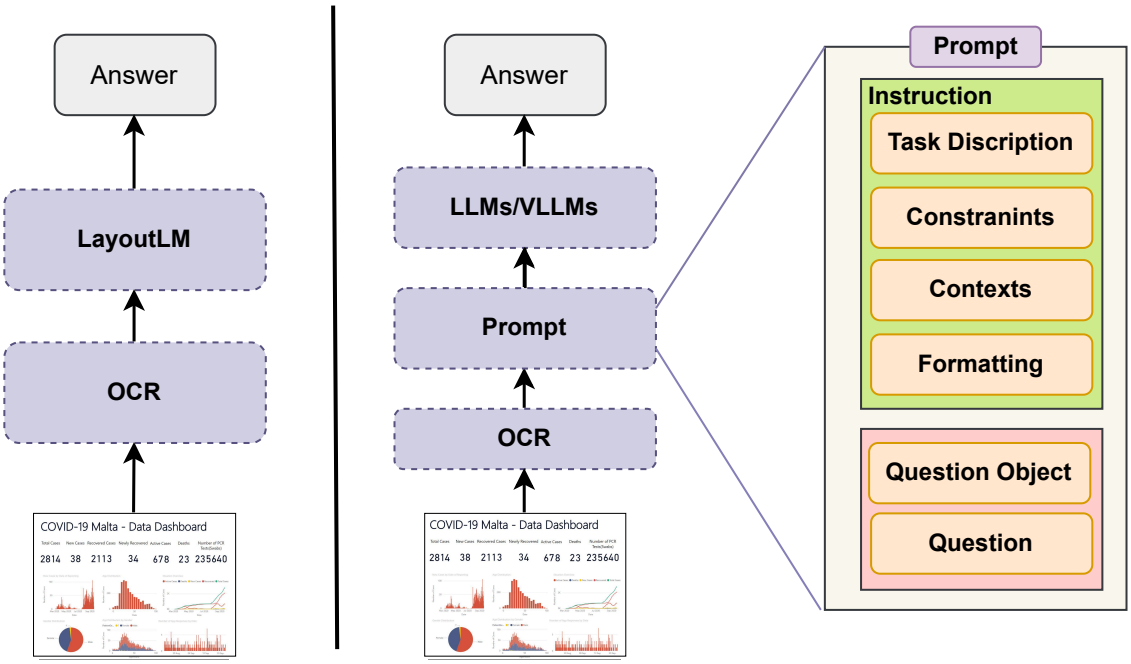


图 4.2 现有多模态模型（左）与大模型（右）在文档理解任务上的对比

4.3 模型架构

图 4.3展示了我们的方法流程，通过多个步骤整合阅读顺序与人类视觉注意力，以增强大型语言模型（LLMs）或视觉语言大模型（VLLMs）对多模态文档的理解能力。为了收集眼动数据，我们使用了 Tobii Pro TX300 眼动仪和 Tobii Studio 软件。受试者阅读的 150 篇文档选自 InfographicVQA[98]数据集。这些文件涵盖了各种格式，从简单的表格到复杂、视觉丰富的文本和图形布局。在对这些文档进行审核后，我们舍弃了其中一些文档，最终得到了 146 份文档。这种多样化的选择有助于全面分析参与者在理解文档时如何整合文本和视觉信息。在阅读过程中记录眼球运动，包括凝视点、注视持续时间、频率、扫视距离和瞳孔大小。纠正了数据异常，解决了缺失或异常的点。通过考虑一致性、相关性和任务性能等因素来评估数据的质量和代表性，我们最终手动选择了最佳注释作为人工参考数据。

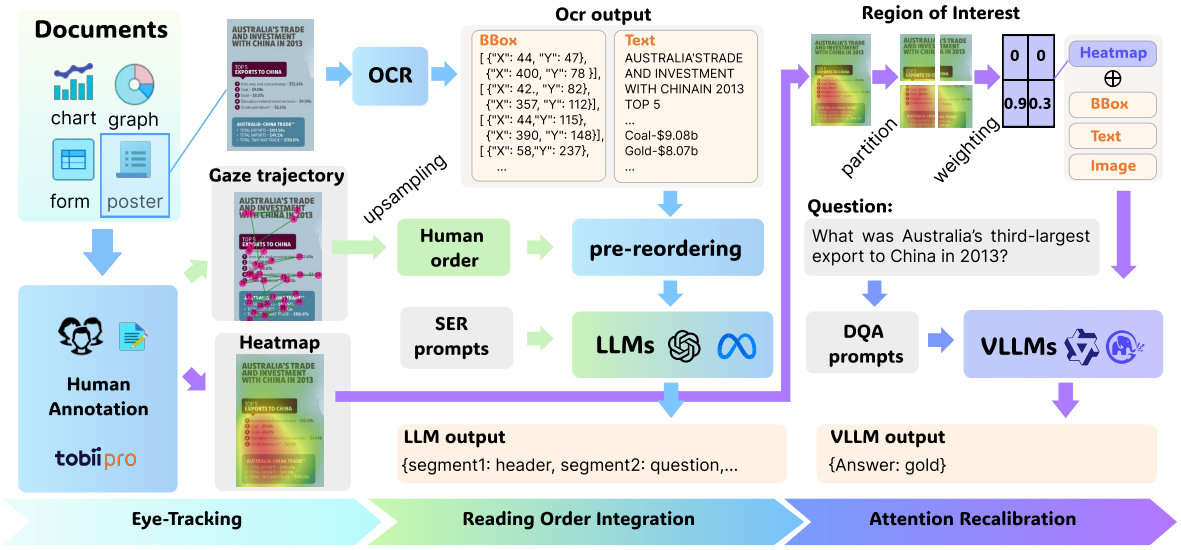


图 4.3 将阅读顺序和视觉热图整合到大型语言模型（LLMs）或视觉语言大型模型（VLLMs）的流程：尽管我们展示了将视觉热图整合到视觉语言大型模型（VLLMs）中的方法，但也可以通过将这些权重与输入的边界框相结合，将其整合到大型语言模型（LLMs）中

具体而言，我们的方法流程包含以下几个关键步骤：（1）首先是文本内容与边界框的提取，利用 OCR 技术从文档中提取文本内容及其对应的边界框坐标，确保模型能够获取文本的位置信息，为后续的阅读顺序整合提供基础。（2）然后是人工标注阅读顺序与任务提示的结合，通过人工标注的阅读顺序（如从左到右、从上到下）和任务提示（如问答模板），为模型提供结构化的引导，帮助模型理解文档的逻辑结

构。眼动仪记录固定点和注视持续时间，然后汇总成热图，其中每个像素的强度反映了固定频率或持续时间。对于 VLLMs，热图直接输入视觉编码器。对于 LLMs，我们将其转换为热矩阵，将其与边界框坐标对齐，并在 JSON 文件中用热值扩展 OCR 边界框条目（见图 4.3）。热矩阵被归一化为 $[0, 1]$ 范围，以便在不同文档中保持一致。为了帮助 LLMs 和 VLLMs 解释数据，我们还添加了一个提示，突出显示最受关注的区域，从而加强视觉和文本信息的整合。例如，高关注区域（如标题、关键数据）的文本会被赋予更高权重，使模型在处理这些区域时投入更多的计算资源。再者是感兴趣区域的提取与内容过滤，针对具体任务（如 “澳大利亚 2013 年对华第三大出口商品”），进一步提取人类感兴趣区域（ROI），过滤无关内容，缩小模型处理范围，提高模型的效率和准确性。最后是模型整合与多模态推理，根据文档类型选择整合方式。对于 VLLMs，直接融合文本、边界框和热图权重，利用视觉-语言联合编码器进行多模态推理；而对于 LLMs，则将热图权重转化为位置编码增强特征，通过调整注意力机制模拟视觉关注，提升模型对视觉信息的理解能力。

4.3.1 基于眼动追踪的热图获取方法

本小节对应这图 4.3中的 Eye-Tracking 部分。眼动追踪实验流程和数据处理与第 3.3.1 节相同，不同的是本次实验我们精心招募了 36 名参与者，其中包括 27 名男性和 9 名女性，他们的平均年龄为 22 岁。这些参与者均为计算机科学或人工智能专业的本科生或研究生，具备扎实的专业知识和技能，能够更好地理解和完成研究任务。此外，所有参与者均为英语作为第二语言（ESL）的学习者，他们的英语流利程度至少相当于雅思 6.0 分、托福 80 分或大学英语六级（CET-6）。这一语言水平确保了他们在研究过程中能够准确理解和表达相关信息，从而保证研究结果的有效性和可靠性。

除此之外，为了确保研究结果的准确性和可靠性，我们让参与者带着问题寻找答案，对参与者的回答进行了严格的准确性检查。在数据收集过程中，任何不正确的答案都会被舍弃，以避免对研究结果产生误导性影响。对于某些关键问题，如果参与者初次回答不准确，我们会对其进行重新测试，直到获得完全正确的答案为止。这种严格的质量控制措施有效提高了数据的准确性和可信度，为后续的研究分析奠定了坚实的基础。

在数据处理阶段，每份文档都经过了至少三处人工注释。这一过程由专业的标注

团队完成，他们根据预定义的标准和指南对文档中的关键信息进行标注和分类。多处人工注释不仅可以提高数据的丰富性和多样性，还可以通过对比不同标注者的标注结果来验证数据的一致性和准确性。尽管由于高标注成本，我们未能实现由多名标注者对所有数据进行标注，但通过至少三处人工注释，我们仍然能够在一定程度上确保数据质量，为后续的研究分析提供可靠的数据支持。最后，我们直接从眼动追踪记录仪中将每张文档的热图导出用于后续的实验，如图 4.1所示。

通过以上措施，我们在参与者招募、培训、数据收集和处理等各个环节都严格把控质量，确保了研究的科学性和有效性。这些努力不仅为本次研究提供了高质量的数据支持，也为相关领域的进一步研究积累了宝贵的经验。

4.3.2 基于热图提示的文档理解方法

(1) 提示构建

在文档问答（Documents Question Answering, DQA）任务中，大模型收到查询并基于结构化的提示生成响应。这种提示机制确保了模型的回答与用户查询高度相关，从而提高了答案的准确性。例如，当用户查询 “What is the official language of Spain?” 时，模型将生成如下响应：{“Answer”: “Spanish”} 如果模型在文档中未找到相关信息，则响应为：{“Answer”: “No”}。这种方法通过将文档内容与用户查询紧密结合，显著提升了模型在视觉富文档（VRDs）理解任务中的性能。值得注意的是，用于格式化的示例在测试数据集中从未出现过，从而避免了模型对特定示例的过拟合，确保了模型的泛化能力。

通过这种结构化的提示机制，模型能够更准确地理解用户查询，并从文档中提取相关信息，从而生成高质量的答案。这种方法不仅提高了模型的性能，还增强了其在复杂文档理解任务中的实用性。

本小节阐述如何为 LLMs 构建有效的提示框架，主要包括指令构建和查询提示构建两个核心部分。图 4.2（右）展示了 LLMs 提示的基本结构框架。

1）任务描述（TaskDescription）：每个任务都需要明确的任务描述 I_t ，以引导 LLMs 专注于特定目标并激活相关知识。任务描述应精心设计，使用清晰的自然语言进行增强，使 LLMs 更深入地理解任务目标。以文档问答任务为例，任务描述可设计为：

“你是一位文档问答专家。输入是一张信息图表文档的 OCR 结果，包括文本及

其位置坐标（文本左上角坐标和文本右下角坐标）。接下来的输入将是一个问题，你需要回答这个问题。例如，问题是：“你多大了？”输出应为：{“Answer”: “12”}。如果你找不到正确答案，输出应为：{“Answer”: “No”}。”

2）标签映射（LabelMapping）：视觉问答任务需要为每个问题预测一个正确答案。因此，指令中需明确 LLMs 的预测标签空间（LabelSpace）。标签映射 I_l 的目标是将语义相似的答案映射到统一的标签空间，使 LLMs 能够有效执行分类任务。为实现这一目标，可从数据集中收集原始答案及其文本描述，例如 “红色” 代表 “图像中物体的颜色”。将原始答案 (Y') 及其描述 Y 包含在上下文中，以提示 LLMs 解决测试样例。标签映射可形式化表示为：

$$I_l = (Y'_1, Y_1), (Y'_2, Y_2), ..., (Y'_n, Y_n) \tag{4.1}$$

3）上下文范例（ContextDemonstrations）：基于上下文学习的思路，可在零样本推理的基础上添加上下文范例作为任务提示。以视觉问答为例，上下文范例通常是一些具有代表性的样本，例如图像中的显著特征、问题与答案的对应关系等。上下文范例通常包含三部分：

图像描述：包括图像内容 I_c 及其关键区域的位置信息 B_c 问题：例如 “图像中的物体是什么颜色?”；正确答案：“红色”。上下文范例并非必须，若包含则 LLMs 进行少样本推理，否则为零样本推理。视觉问答的上下文范例可形式化为：

$$ContextDemonstrations = (I_c, B_c, Q_c, Y_c) \tag{4.2}$$

4）约束条件（Constraints）：在 LLMs 的推理过程中，由于模型能力过于广泛，可能会产生无意义或不合理的结果。因此，在指令中加入适当的约束条件 I_{cons} ，以限制模型生成不符合任务要求的输出。例如，在文档字问答任务中，可以约束模型回答的结果必须是数字且带有单位，因为这种结果会影响最终模型评估的效果。

5）格式化输出（Formatting）：为了确保 LLMs 的输出格式一致且标准化，便于后续的数据处理与评估，指令中需要明确预期的输出格式 I_f 。以视觉富文档（VRDs）信息抽取任务为例，LLMs 的输出应该简洁，通常以易于处理的 JSON 格式保存。格式化输出的示例如下：{“Answer”: “xxx”}。

6）查询提示（Query）：查询提示 Q 的设计与上下文范例类似，但其核心区别在于，查询部分由问题对象（QuestionObject）和具体问题组成，而标签是需要 LLMs

预测的目标。以视觉问答为例，问题对象可以是图像中的某个区域或物体，具体问题则是对该对象的提问。查询提示的形式化表示如下：

$$Q = \{QuestionObject, Question\} \tag{4.3}$$

例如，在视觉问答任务中，问题对象可以是图像中的某个区域，具体问题则是对该区域的提问，如“图像中物体的颜色是什么?”。通过这种方式，LLMs 能够生成结构化的输出，为后续任务提供可靠的支持。

这种提示构建方法不仅简化了模型的应用流程，还提升了输出的规范性和可解释性，为下游任务的高效处理奠定了基础。

7）推理（Inference）：在构建完整的指令和查询提示后，LLMs 能够根据序列化的提示生成具体的预测结果。在文档视觉问答（DQA）任务中，模型需要根据输入的图像和问题生成相应的答案。首先将图像 I 的视觉信息（如图像描述或关键区域位置）与问题 Q 结合，同时融入任务指令 I_{inst} 和查询提示 Q_{query}，形成完整的输入序列。然后 LLMs 根据输入序列，结合预训练知识，逐步推理问题的答案。最后模型生成符合格式化要求 I_f 的输出，包括问题 ID、问题内容、图像描述和预测答案。例如，对于一个问题 Q，其输出形式为：{ “Answer”:“红色”}。这一过程可以形式化表示为：给定图像 I 和相关问题 Q，LLMs 根据指令 I_{inst}（包括任务描述、标签映射、约束条件和格式化输出要求）和查询提示 Q_{query}，生成问题的预测答案。推理过程可表示为：

$$A = \text{LLMs}(I_{inst}, Q_{query}, I, Q) \tag{4.4}$$

其中，A 表示模型生成的答案，I_{inst} 为任务指令，Q_{query} 为查询提示，I 为输入图像，Q 为具体问题。

生成的预测答案可直接用于下游任务，如智能问答系统、图像检索或人机交互场景。由于输出格式标准化，后续处理过程更加高效且易于集成到实际应用中。通过这种推理机制，LLMs 能够在无需额外微调的情况下，高效完成视觉问答任务，充分展现了其强大的泛化能力和多模态理解能力。

4.3.3 基于 HSV 色彩空间的感兴趣区域划分方法

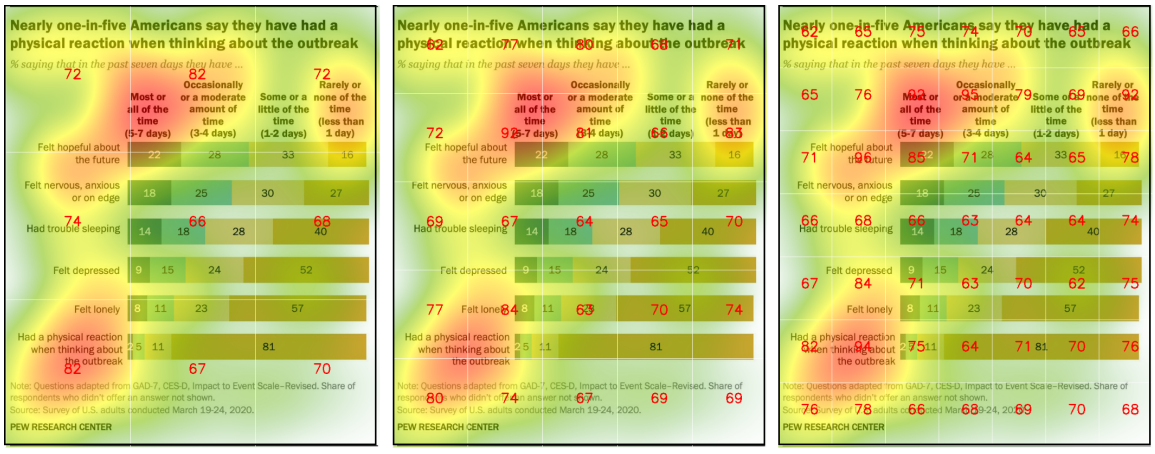
本小节对应图 4.3中 Attention Recalibration 上方内容。

（1）相似性区域划分方法

在文档视觉问答（DQA）任务中，人类阅读者通常会将更多的注意力集中在答案可能出现的区域。区域越显著（在热图中以红色区域表示），人类在寻找答案时投入的认知努力就越多。为了检验感兴趣区域（ROI）对 DQA 任务的影响，我们计算了三种不同网格粒度下的平均显著性得分：3×3、5×5 和 7×7 区域。对于每个区域，我们首先计算热图与 HSV 颜色空间中原始图像之间的颜色差异。基于这种颜色差异，我们为每个区域分配一个显著性得分，其中绿色、黄色和红色区域分别对应逐渐增加的显著性得分（绿色：40-60，黄色：60-80，红色：80-100）。得分低于 40 的区域被认为注意力负载较低。使用每个文档的整体平均显著性得分，我们然后将 InfoGraph 数据集分为三个子集：低、中、高。

具体来说，我们首先对每个文档进行网格划分，分别采用 3×3、5×5 和 7×7 的网格粒度，以覆盖文档的不同区域。对于每个网格区域，我们计算热图与原始图像在 HSV 颜色空间中的颜色差异。HSV 颜色空间是一种常用的色彩模型，能够更好地反映人类对颜色的感知。颜色差异的计算有助于我们确定哪些区域在热图中更为显著，即更受人类关注。

接下来，根据计算出的颜色差异，我们为每个区域分配一个显著性得分。绿色区域表示显著性较低，得分在 40-60 之间；黄色区域表示中等显著性，得分在 60-80 之间；红色区域表示高显著性，得分在 80-100 之间。得分低于 40 的区域则被认为是注意力负载较低的区域，可能对 DQA 任务的贡献较小，如图 4.4所示。最后，我们



(a) (b) (c)

图 4.4 热力区域划分结果：3×3 (a)，5×5 (b) 和 7×7 (c)

计算每个文档的整体平均显著性得分，并根据得分将 InfoGraph 数据集划分为三个子集：低显著性、中显著性和高显著性。这种划分有助于我们进一步分析不同显著性水平的文档对 DQA 任务性能的影响，从而更好地理解人类注意力分配在文档问答中的作用。不同的色彩空间能够从不同角度反映图像的特性，为注意力在图像中的分配提供可视化。在本研究中，我们将原始图像和热力图图像从 RGB 色彩空间转换为 HSV 色彩空间，其目的是为了更好地分析图像的颜色和亮度信息。可以形式化表述为以下过程：

1）图像预处理：首先，我们将原始图像和热力图图像从 RGB 色彩空间转换为 HSV 色彩空间。HSV 色彩空间由色相（Hue）、饱和度（Saturation）和亮度（Value）三个通道组成，能够更好地反映图像的颜色和亮度信息。此转换基于人眼对色相变化的敏感性高于亮度变化的视觉特性。转换公式如下：

$$H_{\text{hsv}} = \text{cvtColor}(I_{\text{heatmap}}, \text{COLOR\_BGR2HSV}) \tag{4.5}$$

$$O_{\text{hsv}} = \text{cvtColor}(I_{\text{original}}, \text{COLOR\_BGR2HSV}) \tag{4.6}$$

其中：

- I_{heatmap}：热力图图像，表示注意力权重分布的概率图；
- I_{original}：原始文档图像，分辨率为 $W \times H$ 像素；
- 选择 BGR 转换而非 RGB，因 OpenCV 默认图像读取格式为 BGR 通道顺序；
- HSV 各通道范围：H$\in [0, 179]$, S$\in [0, 255]$, V$\in [0, 255]$（8 位量化）。

2）提取 HSV 通道：在 HSV 色彩空间中，我们分别提取热力图和原始图像的色相（H）、饱和度（S）和亮度（V）通道。此过程保留了颜色空间的解耦特性，便于后续独立分析各视觉特征：

$$H_h = H_{\text{hsv}}[:, :, 0] \tag{4.7}$$

$$H_s = H_{\text{hsv}}[:, :, 1] \tag{4.8}$$

$$H_v = H_{\text{hsv}}[:, :, 2] \tag{4.9}$$

$$O_h = O_{\text{hsv}}[:, :, 0] \tag{4.10}$$

$$O_s = O_{\text{hsv}}[:, :, 1] \tag{4.11}$$

$$O_v = O_{\text{hsv}}[:, :, 2] \tag{4.12}$$

其中：

- H_h, H_s, H_v：热力图在 HSV 空间的三个特征通道；
- O_h, O_s, O_v：原始图像对应的特征通道；
- 通道索引 0/1/2 分别对应 H/S/V 通道。

3）计算通道差异：计算热力图与原始图像在各通道上的绝对差异，采用绝对值运算而非平方差以保持差异值的线性特性：

$$\Delta_h = |H_h - O_h| \tag{4.13}$$

$$\Delta_s = |H_s - O_s| \tag{4.14}$$

$$\Delta_v = |H_v - O_v| \tag{4.15}$$

其中：

- Δ_h：色相差异矩阵，反映注意力区域的颜色偏移；
- Δ_s：饱和度差异矩阵，表征关注区域的色彩强度变化；
- Δ_v：亮度差异矩阵，描述明暗对比变化；
- 所有差异矩阵维度保持为 $H \times W$，与原始图像分辨率一致。

4）归一化差异：对每个通道进行最小-最大归一化，消除量纲差异并统一到 [0,1] 区间：

$$\Delta_h^{\text{norm}} = \frac{\Delta_h - \min(\Delta_h)}{\max(\Delta_h) - \min(\Delta_h)} \tag{4.16}$$

$$\Delta_s^{\text{norm}} = \frac{\Delta_s - \min(\Delta_s)}{\max(\Delta_s) - \min(\Delta_s)} \tag{4.17}$$

$$\Delta_v^{\text{norm}} = \frac{\Delta_v - \min(\Delta_v)}{\max(\Delta_v) - \min(\Delta_v)} \tag{4.18}$$

归一化处理使得：完全无差异区域得分为 0，最大差异区域得分为 1，并且各通道差异具有可比性。

5）计算总差异：通过加权线性组合生成综合差异度量，权重分配基于特征重要性分析：

$$\Delta_{\text{total}} = \alpha\Delta_h^{\text{norm}} + \beta\Delta_s^{\text{norm}} + \gamma\Delta_v^{\text{norm}} \tag{4.19}$$

权重配置说明：饱和度权重最高（2.4），因人类视觉对色彩纯度变化敏感；色相权重中等（0.5），反映颜色本质特征的改变；亮度权重最低（0.1），降低光照条件干扰。

6）多级颜色阈值划分：根据总差异值定义四级颜色显著性区域：

$$\text{ColorLevel} = \begin{cases} \text{红色} & \Delta_{\text{total}} \in (0.7, 1.0] \\ \text{黄色} & \Delta_{\text{total}} \in (0.5, 0.7] \\ \text{绿色} & \Delta_{\text{total}} \in (0.3, 0.5] \\ \text{背景} & \Delta_{\text{total}} \in [0.0, 0.3] \end{cases} \tag{4.20}$$

阈值设定依据：当红色阈值 0.7：对应人类视觉对红色敏感度的 JND(Just Noticeable Difference)[108] 理论值；颜色区间宽度 0.2 保证足够区分度。

7）动态颜色评分机制：对每个网格单元 (i, j) 计算颜色分布：

$$P_{i,j}^{\text{red}} = \frac{\sum(\text{ColorLevel} = \text{红色})}{w_{\text{cell}} \times h_{\text{cell}}} \tag{4.21}$$

$$P_{i,j}^{\text{yellow}} = \frac{\sum(\text{ColorLevel} = \text{黄色})}{w_{\text{cell}} \times h_{\text{cell}}} \tag{4.22}$$

$$P_{i,j}^{\text{green}} = \frac{\sum(\text{ColorLevel} = \text{绿色})}{w_{\text{cell}} \times h_{\text{cell}}} \tag{4.23}$$

采用分段评分函数：

$$S_{i,j} = \begin{cases} 80 + 20 \tanh(5P_{i,j}^{\text{red}}) & P_{i,j}^{\text{red}} > 0.25 \\ 60 + 20 \sin(\frac{\pi}{2} P_{i,j}^{\text{yellow}}) & P_{i,j}^{\text{yellow}} > 0.35 \\ 40 + 20\sqrt{P_{i,j}^{\text{green}}} & P_{i,j}^{\text{green}} > 0.45 \\ 40 P_{i,j}^{\text{other}} & \text{其他} \end{cases} \tag{4.24}$$

函数特性：

- 红色区域：tanh 函数快速饱和，保证 $P > 0.4$ 时 $S > 95$；
- 黄色区域：正弦函数平滑过渡，$P = 0.5$ 时 $S = 70$；
- 绿色区域：平方根函数缓变增长，$P = 1.0$ 时 $S = 60$；
- 背景区域：线性映射，$S \leq 40$。

8）文档级加权聚合：定义颜色权重矩阵：

$$w_{i,j} = \begin{cases} 3.0 & \text{红色网格} \\ 2.0 & \text{黄色网格} \\ 1.5 & \text{绿色网格} \\ 1.0 & \text{背景网格} \end{cases} \tag{4.25}$$

计算文档显著性得分：

$$S_{\text{doc}} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} w_{i,j} S_{i,j}}{\sum_{i=1}^{R}\sum_{j=1}^{C} w_{i,j}} \tag{4.26}$$

（2）区域相似性评估方法

为了解决大语言模型在每个区域的注意力相似性的评估问题，本文提出一种融合视觉注意力机制的空间推理评估框架，通过结构化区域显著性分数的量化与大模型提示工程的协同作用，实现文档问答任务中答案定位能力的精细化测评。该方法实施过程包括三个主要阶段：

1）文档空间离散化与显著性建模。采用多粒度网格划分策略（分别为 3×3、5×5、7×7 单元），将文档图像划分为规则的区域。根据 HSV 色彩空间转换技术，计算每个单元区域的色相、饱和度及亮度差异值，通过归一化处理与加权融合生成区域显著性评分 $S_{grid} \in [0, 100]$。依据人类视觉感知特性建立三级显著性分类标准：高显著性红色区域（$S_{grid} \geq 80$）、中显著性黄色区域（$60 \leq S_{grid} < 80$）、低显著性绿色区域（$40 \leq S_{grid} < 60$）。最终构建包含区域坐标边界、唯一标识符 ID 的显著性评分的标准化 JSON 元数据文件，形成机器可解析的视觉注意力分数矩阵。

2）空间感知的多模态提示工程。设计双层级指令架构引导视觉语言模型执行区域定位任务。系统级指令嵌入空间认知先验，声明“候选区域的显著性评分反映人类视觉关注度分布”；任务级指令动态注入结构化区域信息，明确要求大模型根据文档内容与显著性分布输出最可能的答案区域唯一标识符 ID。通过约束输出格式确保结果可程序化解析，同时采用显著性排序策略优化模型的注意力分配机制。具体来说，我们将 4.3.1 中划分的区域的坐标，序号和分数保存到 json 文件中，同时给大模型输入 json 和 Prompt 提示，提示如下：

“你是一位文档问答专家。输入是一张信息图表文档的多个区域的坐标，序号 id

和区域分数（权重）。接下来的输入将是一个问题，你需要回答这个问题，答案也可能出现在权重较高的区域。例如，问题是："你多大了？"输出应为：{"Answer": "0"}。如果你找不到正确答案，输出应为：{"Answer": "No"}。"

3）答案区域定位准确率计算。定义基于显著性权重的评估指标，公式如下：

$$\text{ARLA} = \frac{S_{pred}}{S_{average}} \times 100\% \tag{4.27}$$

其中 S_{pred} 表示模型预测区域的显著性评分，$S_{average}$ 为人工标注真实答案区域的平均显著性值。该指标通过量化预测区域与真实区域的显著性关联强度，实现连续空间误差的细粒度表征。当预测区域与真实区域完全匹配时获得满分，若存在定位偏差则依据显著性衰减比例反映准确率损失程度，突破传统二元评估范式对局部误差不敏感的局限性。

4.4 实验与分析

4.4.1 评价指标

在视觉问答（DQA）任务中，平均归一化 Levenshtein 相似度[109]（Average Normalized Levenshtein Similarity，ANLS）是一种广泛使用的评价指标。它主要用于评估生成答案与标准答案之间的匹配程度，尤其适用于处理 OCR 识别错误或生成模型输出中的文本偏差。ANLS 的核心优势在于其能够有效应对文本生成任务中的不确定性，具体特点如下：容错性强，例如 ANLS 基于编辑距离（Levenshtein 距离），能够容忍拼写错误、多余字符或轻微词汇差异，从而更灵活地评估生成答案的准确性；归一化设计，例如通过将编辑距离除以文本的最大长度，ANLS 实现了对不同长度文本的公平比较，避免了长文本偏差问题；适用性广，如 ANLS 不仅适用于视觉问答任务，还可用于其他如文档问答或文本生成任务等需要生成自然语言答案的场景；鲁棒性高，例如在面对 OCR 识别错误或生成模型输出不完美的情况下，ANLS 能够提供更为合理的评分，减少因微小差异导致的评分偏差。其公式如下：

$$\text{ANLS} = \frac{1}{N} \sum_{i=1}^{N} \text{NLS}(A_i, P_i) \tag{4.28}$$

其中：

- N 是样本总数；

- A_i 第 i 个样本的标准答案；
- P_i 第 i 个样本的标准答案；
- $\text{NLS}(A_i, P_i)$ 是归一化 Levenshtein 相似度，计算公式为：

$$\text{NLS}(A, P) = 1 - \frac{\text{Levenshtein}(A, P)}{\max(|A|, |P|)} \tag{4.29}$$

4.4.2 实验环境和细节

本章实验所使用的软硬件环境如表 4.1所示。

表 4.1 实验的软硬件环境

部件	参数
操作系统	Ubuntu 20.04.2
系统内存	976G
CPU 处理器	Intel(R) Xeon(R) Gold 6348 CPU @2.60GHz
GPU 处理器	A800-PCIE-80GB
Python 版本	2.3.0
Pytorch 版本	3.10
CUDA 版本	12.1
transformer 版本	4.5.0

我们使用的 LLaMA3 8B、GPT-3.5、GPT-4.0 是纯语言大模型，GLM-4V 9B 和 Qwen2-VL 7B 视觉语言大模型，它们的具体参数如表 4.2所示。

表 4.2 LLMs/VLLMs 参数列表

模型名称	发布时间	参数规模	训练数据量	上下文长度	语言	是否开源
LLaMA3 8B	2023-03	8B	15 万亿个 Token	8K	英语	是
GPT-3.5	2022-11	175B	近 1 万亿单词	16K	中、英	否
GPT-4.0	2023-03	未知	近 13 万亿单词	8K	中、英	否
GLM-4V 9B	2024-06	9B	10 万亿图文对	8K	中、英	是
Qwen2-VL 7B	2024-09	7B	1.4 万亿个图文对	128K	中、英	是

4.4.3 实验结果

（1）大型语言模型（LLMs）的视觉注意力是否与人类的一致？

我们进一步评估了大模型与人类注意力机制的相关性，我们在不同显著性的子集上验证的结果如图 4.5所示，根据图中的实验结果，我们分析了 GLM-4V-9B 和 Qwen2-VL-7B 两个模型在不同显著性子集（低、中、高）和网格粒度（3×3、5×5、7×7）下的答案区域定位准确率（ARLA）表现。首先，显著性水平对模型的 ARLA 有显著影响。高显著性数据集在三种不同的网格粒度上的 ARLA 值均显著高于中、低显著性数据集。例如 Qwen2-VL-7B 在高显著性的 3×3 网格条件下的 ARLA 达到了 93.7%，而相同条件下中低显著性数据集分的 ARLA 分别为为 88.3% 和 86.9%。这表明显著性信息能够有效引导大语言模型对答案区域的定位，缩小搜索空间。相反，在低显著性数据集上，由于缺乏明确的注意力线索，模型容易受到背景干扰，导致 ARLA 表现不佳。

网格粒度对定位精度的影响也十分显著。随着网格粒度从 3×3 增至 7×7，ARLA 值呈现下降趋势。例如，Qwen2-VL-7B 在高显著性条件下，ARLA 从 3×3 网格的 93.7% 下降到 7×7 网格的 87.9%。这表明粗粒度划分的区域更大能够捕捉更广泛的空间特征。

总而言之，显著性与网格粒度具有协同效应，高显著性与粗粒度网格的组合是模型达到峰值性能的最优条件。这一发现证明了大模型利用这种显著性信息可以更好的理解人类视觉注意力机制，实现与人类的视觉注意力机制有很强的相关性。

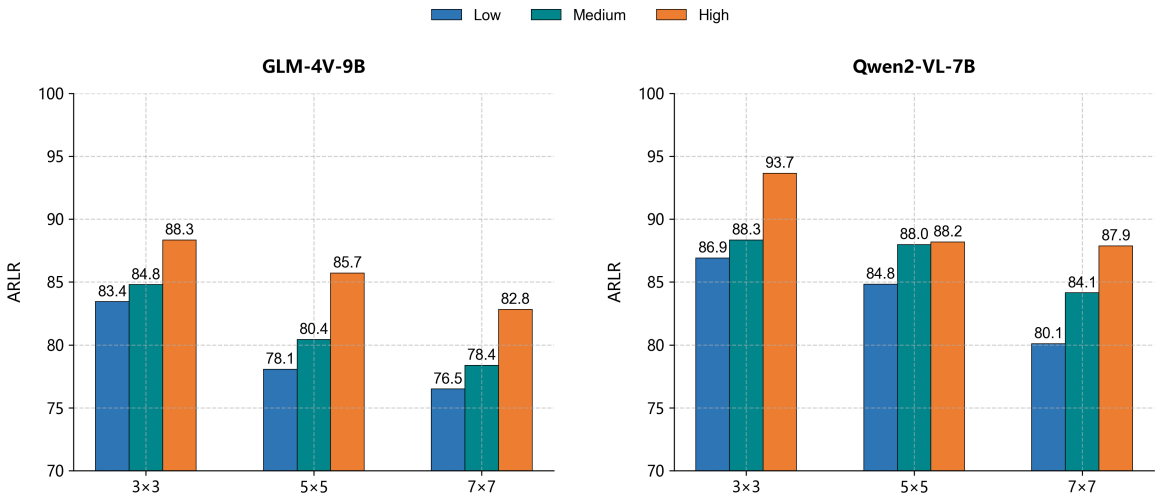


图 4.5　不同显著性子数据集（低、中、高）在不同网格粒度下的答案区域定位准确率（ARLA）

（2）人类的视觉上注意力是否对视觉语言大型模型（VLLMs）的文档视觉问答（DQA）任务有积极影响？

如表 4.3所示，随着关注点转移到更高显著性的区域，GLM-4V-9B 和 Qwen2-VL-7B 模型的性能均有所提升，在高显著性子集中得分最高。随着显著性的增加，ARLA（答案区域定位准确率）、F1 分数和 ANLS（归一化答案位置相似度）等性能指标均有所提高，这表明模型在人类注意力更高的区域定位答案更为准确。具体而言，当关注点集中在高显著性区域时，模型能够更有效地识别和定位答案，从而提高整体性能。GLM-4V-9B 和 Qwen2-VL-7B 在高显著性子集中的表现优于低显著性和中显著性子集，这表明显著性区域对模型性能有显著影响。

在模型性能方面，Qwen2-VL-7B 在所有显著性数据集上的表现均优于 GLM-4V-9B。例如，在高显著性数据集上，Qwen2-VL-7B 的 ARLA 为 89.90%，比 GLM-4V-9B 的 85.62% 高出 4.28 个百分点；F1 分数为 42.59%，比 GLM-4V-9B 的 43.01% 略低，但在 ANLS 上表现更好，达到 63.94%，比 GLM-4V-9B 的 62.32% 高出 1.62 个百分点。Qwen2-VL-7B 在中等和高显著性子集中的 ARLA、F1 和 ANLS 指标上均优于 GLM-4V-9B，这可能是因为 Qwen2-VL-7B 在处理视觉和语言信息时具有更强的集成能力和更精细的注意力机制，使其能够更准确地定位答案区域。

通过 ROI 分割利用人类注意力模式，可以显著增强模型在 DQA 任务中的表现。这种方法不仅提高了模型在高显著性区域的性能，还展示了人类注意力数据在优化视觉语言模型方面的潜力。未来的研究可以进一步探索如何更有效地整合人类注意力模式，以提升模型在各种视觉语言任务中的性能。总体而言，结果证实了通过 ROI 分割利用人类注意力模式，能够显著提升模型在 DQA 任务中的性能。

表 4.3　在具有显著性（Saliency）和答案区域定位准确率（ARLA）得分的数据集上比较视觉语言大型模型（VLLMs）的性能

Model	**Dataset**	**Saliency**(%)↑	**ARLA**(%)↑	**F1**(%)↑	**ANLS**(%)↑
GLM-4V 9B	Low	73.06	79.33	41.48	59.66
	Medium	84.61	81.20	41.89	60.50
	High	**91.29**	**85.62**	**43.01**	**62.32**
Qwen2-VL 7B	Low	73.06	83.94	39.77	62.82
	Medium	84.61	86.82	40.16	63.05
	High	**91.29**	**89.90**	**42.59**	**63.94**

（3）人类认知如何增强大型语言模型？

表 4.4展示了不同大型语言模型（LLMs）和视觉语言大型模型（VLLMs）在具

有不同模态和配置的文档问答（DQA）任务中的性能结果，重点关注了阅读顺序与注意力重校准的不同策略之间的协同作用。研究发现，引入注意力重校准机制以及 Z-order 策略通常能够提升大多数模型在 DQA 任务中的性能，而最佳结果往往是在这两种方法同时应用时取得的。在这些模型中，Qwen2-VL-7B 和 GPT-4.0 表现出最一致且显著的性能提升。具体而言，Qwen2-VL-7B 在 F1 分数和 ANLS（归一化答案位置相似度）两项指标上均取得了最高分（F1 = 44.73，ANLS = 67.39），而 GPT-4.0 在应用所有技术时也显示出显著的改进。尽管 LLaMA3-8B 和 GPT-3.5 在经过重校准和 Z-order 调整后性能有所提升，但其提升幅度相对较小。这些结果表明，像 Qwen2-VL-7B 和 GPT-4.0 这样的模型特别适合利用注意力重校准和顺序调整，从而在视觉问答任务中显著提升其性能表现。

表 4.4　探索阅读顺序与注意力重校准相结合对文档问答（DQA）任务的影响

Model	**Modalities**		**Order**	**HeatMap**	**F1** (%) ↑	**ANLS** (%) ↑
	BBox+Text	Image	+Z-order			
LLaMA3 8B	✓				18.51	18.51
	✓			✓	**20.64** ↑	**23.91** ↑
	✓		✓		22.65	31.67
	✓		✓	✓	**22.78** ↑	**32.45** ↑
GPT-3.5	✓				22.01	26.45
	✓			✓	**23.17** ↑	**28.89** ↑
	✓		✓		28.43	33.33
	✓		✓	✓	**27.19** ↓	**33.82** ↑
GPT-4.0	✓				31.08	38.76
	✓			✓	**32.67** ↑	**40.03** ↑
	✓		✓		37.49	44.53
	✓		✓	✓	**37.66** ↑	**45.21** ↑
GLM-4V 9B		✓			42.18	61.70
		✓		✓	**43.32** ↑	**64.42** ↑
	✓	✓			43.52	61.30
	✓	✓		✓	44.51	65.41
	✓	✓	✓		43.10	65.04
	✓	✓	✓	✓	**47.25** ↑	**66.30** ↑
Qwen2-VL 7B		✓			41.06	63.74
		✓		✓	**43.53** ↑	**65.38** ↑
	✓	✓			40.95	62.81
	✓	✓		✓	42.23	65.91
	✓	✓	✓		42.50	66.29
	✓	✓	✓	✓	**44.73** ↑	**67.39** ↑

4.5 本章小结

本章主要从 LLMs 和 VLLMs 在视觉富文档理解任务的角度出发，针对带有注意力热图的视觉富文档通过提示大语言模型感知人类的注意力机制。我们通过设计将文档划分成不同数量区域，根据热图颜色对区域评分，然后让 LLMs 和 VLLMs 去选择答案可能所在的区域得到大模型与人力注意力的相似性。并在下游任务的结果中证明了大模型与人类注意力相似性越高，在下游 DQA 任务中结果越好，证明了我们的方法有效行。

第五章　总结和展望

5.1　总结

本文在信息化时代快速发展的背景下，针对视觉富文档（VRDs）在多个行业中的深入应用，探讨了如何通过模拟人类的认知机制来提升人工智能模型对这类文档的理解能力。视觉富文档不仅包含文本信息，还通过版面设计和视觉元素传递关键语义信息，其布局结构能够揭示文本内容之间的内在联系。然而，由于视觉富文档的复杂版式布局和多样化结构特征，现有的智能模型难以全面实现对这些文档的深度理解。相比之下，人类读者能够凭借阅读顺序的把握和视觉注意力的分配，高效地从视觉富文档中提取所需信息。因此，开发能够模拟人类阅读和理解能力的人工智能模型成为迫切需求。

本文主要围绕人类阅读文档时的顺序和注意力机制在视觉富文档理解中的应用，本文的主要工作和贡献包括以下几个方面：

贡献一：提出基于预排序模型的仿人类阅读顺序嵌入，解决了现有文档理解模型输入序列建模难题。

为了解决现有模型在处理视觉富文档时忽略文本顺序问题以及多模态排序对齐问题，本文首先构建了一个包含人类眼动信息的基准数据集，通过眼动追踪记录仪采集人类阅读文档时的眼动生理信息，并将其中重要的特征（如阅读顺序）提取出来，并与原始文档数据集成。此外，本文还提出了模仿人类阅读顺序的预排序模型，通过文本、坐标、文本 + 坐标和文本 + 坐标 + 图像四种类型的模态来模仿人类的阅读顺序，并将模仿出来的阅读顺序和人类阅读顺序进行相似性评估。实验结果表明，所提出的方法在下游任务上的结果优于原始的 OCR 顺序。

贡献二：提出基于热图提示的仿人类注意力机制学习，解决了大语言模型对文档重点区域关注不足难题。

为了解决大模型在视觉富文档理解任务中缺乏对重点区域关注这一问题，本文提出通过整合人类注意力区域增强大语言模型对文档重点区域的关注能力。通过从 InfographicVQA 数据集中选取多模态文档，并利用眼动仪记录受试者的视觉行为数据，构建了高质量标注基准。研究结果表明，视觉注意力与任务表现存在强关联性，

人类在特定任务中的高准确率与其对高价值区域的聚焦高度相关；同时，AI 模型的注意力对齐效应显著，当模型通过热图权重增强对高关注区域的处理时，其任务表现显著提升。

5.2 展望

本文针对视觉富文档理解任务的实体语义识别和文档问答任务进行了研究并提出了相应的文档智能模型和评估方案，但依旧存在诸多不足和可以改进之处：

（1）更适合大模型的阅读顺序自主探索方法

在本文的工作中，我们通过四种不同的模态模仿了人类阅读顺序，但是尚未考虑到更适合大模型可以自主探索的阅读顺序。我们虽然加入了人类阅读时的眼动顺序，但是没有考虑更多的眼动特征。在未来的工作中，探索更适合大模型的阅读顺序自主探索方法方面，有望取得显著进展。随着研究深入，我们期待加入更多的人类眼动特征（如回视、注视时长等），使大模型将能更精准地模拟人类阅读顺序，自主探索能力将不断提升，能处理更复杂的文档，理解更深层次语义。同时，该方法会结合多模态信息处理，让模型在阅读时整合文本与图像等信息，全面理解文档内容及逻辑。

（2）更适合大模型的高效注意力注入方法

本文提出基于热图提示的仿人类注意力机制学习的方法。首先，注意力相似性评估方法依赖于区域化颜色编码和人工标注，对于大模型无法自主的在文档理解上获取更高效的注意力。其次，我们的数据集数量和种类有限，无法更广泛的应用在不同领域上。未来，我们考虑使用更精准方式整合人类视觉注意力机制数据，让大模型更敏锐地捕捉文档中的关键信息。这不仅能提升模型的任务表现，还能使其更契合人类的认知习惯。同时，我们也将组织更多的人员参与眼动追踪实验，构造一个更大的数据集。

未来，我们期待这些方法能推动文档智能模型更贴近人类认知水平，为复杂文档处理提供更高效、精准的解决方案。同时，鼓励更多研究探索人类认知与 AI 技术的深度融合，以期在文档智能领域取得更大突破。

插图索引

表格索引

参考文献

[1] 崔磊. 文档智能: 数据集、模型和应用[J]. 中文信息学报, 2022, 36(6): 1-19.

[2] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M/OL]. MIT Press, 2016. http://www.deeplearningbook.org.

[3] ZHONG X, SHAFIEIBAVANI E, YEPES A J. Image-based table recognition: data, model, and evaluation[A/OL]. 2020. arXiv: 1911.10683. https://arxiv.org/abs/1911.10683.

[4] PAN S J, YANG Q. A survey on transfer learning[J/OL]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359. DOI: 10.1109/TKDE.2009.191.

[5] XU Y, LI M, CUI L, et al. Layoutlm: Pre-training of text and layout for document image understanding[C/OL]//GUPTA R, LIU Y, TANG J, et al. KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020. ACM, 2020: 1192-1200. https://doi.org/10.1145/3394486.3403172.

[6] XU Y, XU Y, LV T, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding[A/OL]. 2022. arXiv: 2012.14740. https://arxiv.org/abs/2012.14740.

[7] HUANG Y, LV T, CUI L, et al. Layoutlmv3: Pre-training for document AI with unified text and image masking[C/OL]//MAGALHÃES J, BIMBO A D, SATOH S, et al. MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022. ACM, 2022: 4083-4091. https://doi.org/10.1145/3503161.3548112.

[8] LI Y, QIAN Y, YU Y, et al. Structext: Structured text understanding with multi-modal transformers[C/OL]//SHEN H T, ZHUANG Y, SMITH J R, et al. MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021. ACM, 2021: 1912-1920. https://doi.org/10.1145/3474085.3475345.

[9] ZHAI M, LI Y, QIN X, et al. Fast-structext: An efficient hourglass transformer with modality-guided dynamic token merge for document understanding[C/OL]// Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China. ijcai.org, 2023: 5269-5277. https://doi.org/10.24963/ijcai.2023/585. DOI: 10.24963/IJCAI.2023/585.

[10] LYU P, LI Y, ZHOU H, et al. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond[J/OL]. CoRR, 2024, abs/2405.21013. https://doi.org/10.48550/arXiv.2405.21013. DOI: 10.48550/ARXIV.2405.21013.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C/OL]// GUYON I, VON LUXBURG U, BENGIO S, et al. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. 2017: 5998-6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[12] WANG H, CHEN X, WANG R, et al. Vision-enhanced semantic entity recognition in document images via visually-asymmetric consistency learning[C/OL]//BOUAMOR H, PINO J, BALI K. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 15718-15731. https://aclanthology.org/2023.emnlp-main.973/. DOI: 10.18653/v1/2023.emnlp-main.973.

[13] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: A survey[J/OL]. Computational Visual Media, 2022, 8(3): 331–368. http://dx.doi.org/10.1007/s41095-022-0271-y.

[14] SASSIOUI A, BENOUINI R, OUARGUI Y, et al. Visually-rich document understanding: Concepts, taxonomy and challenges[C/OL]//2023: 1-7. DOI: 10.1109/WINCOM59760.2023.10322990.

[15] DING Y, HAN S C, LI Z, et al. David: Domain adaptive visually-rich document understanding with synthetic insights[A/OL]. 2024. arXiv: 2410.01609. https://arxiv.org/abs/2410.01609.

[16] LI Q, LI Z, CAI X, et al. Enhancing visually-rich document understanding via layout structure modeling[C/OL]//MM '23: Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2023: 4513–4523. https://doi.org/10.1145/3581783.3612327.

[17] 章倩, 王梓祺. 基于自定义模板的 OCR 技术及应用[J/OL]. 指挥信息系统与技术, 2023, 14(05): 94-98. DOI: 10.15908/j.cnki.cist.2023.05.015.

[18] CHEN X, JIN L, ZHU Y, et al. Text recognition in the wild: A survey[A/OL]. 2020. arXiv: 2005.03492. https://arxiv.org/abs/2005.03492.

[19] 邱立可, 王晓年, 朱劲, 等. 基于模板匹配和 Tesseract 的票据归类和索引[J]. 计算机与现代化, 2010(07): 132-135.

[20] 张志华. 机器学习的发展历程及启示[J]. 中国计算机学会通讯, 2016, 12(11): 55-60.

[21] ROSENBLATT M. Remarks on some nonparametric estimates of a density function [J/OL]. The Annals of Mathematical Statistics, 1956, 27(3): 832-837[2025-03-24]. http://www.jstor.org/stable/2237390.

[22] PARZEN E. On estimation of a probability density function and mode[J/OL]. Annals of Mathematical Statistics, 1962, 33: 1065-1076. https://api.semanticscholar.org/CorpusID:122932724.

[23] OTSU N. A threshold selection method from gray-level histograms[J/OL]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62-66. DOI: 10.1109/TSMC.1979.4310076.

[24] HARALICK R. Statistical and structural approaches to texture[J/OL]. Proceedings of the IEEE, 1979, 67(5): 786-804. DOI: 10.1109/PROC.1979.11328.

[25] TUCERYAN M, JAIN A K. Texture analysis[M/OL]. 207-248. https://worldscientific.com/doi/abs/10.1142/9789812384737_0007.

[26] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20 (3): 273-297.

[27] RABINER L. A tutorial on hidden markov models and selected applications in speech recognition[J/OL]. Proceedings of the IEEE, 1989, 77(2): 257-286. DOI: 10.1109/5.18626.

[28] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001: 282–289.

[29] KIM Y. Convolutional neural networks for sentence classification[C/OL]// MOSCHITTI A, PANG B, DAELEMANS W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1746-1751. https://aclanthology.org/D14-1181/. DOI: 10.3115/v1/D14-1181.

[30] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[A/OL]. 2015. arXiv: 1409.2329. https://arxiv.org/abs/1409.2329.

[31] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[A/OL]. 2015. arXiv: 1409.1556. https://arxiv.org/abs/1409.1556.

[32] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J/OL]. Neural Computation, 1997, 9(8): 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.

[33] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C/OL]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 6645-6649. DOI: 10.1109/ICASSP.2013.6638947.

[34] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C/OL]//BURSTEIN J, DORAN C, SOLORIO T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186. https://aclanthology.org/N19-1423/. DOI: 10.18653/v1/N19-1423.

[35] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[C/OL]//2018. https://api.semanticscholar.org/CorpusID:49313245.

[36] [C].

[37] WANG Z, XU Y, CUI L. Layoutreader: Pre-training of text and layout for reading order detection[A/OL]. 2021. arXiv: 2108.11591. https://arxiv.org/abs/2108.11591.

[38] GU Z, MENG C, WANG K, et al. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding[C/OL]//2022: 4573-4582. DOI: 10.1109/CVPR52688.2022.00454.

[39] XIE S, GIRSHICK R, DOLLáR P, et al. Aggregated residual transformations for deep neural networks[A/OL]. 2017. arXiv: 1611.05431. https://arxiv.org/abs/1611.05431.

[40] XU Y, LV T, CUI L, et al. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding[A/OL]. 2021. arXiv: 2104.08836. https://arxiv.org/abs/2104.08836.

[41] SYLWESTER D, SETH S. A trainable, single-pass algorithm for column segmentation[C/OL]//Proceedings of 3rd International Conference on Document Analysis and Recognition: Vol. 2. 1995: 615-618 vol.2. DOI: 10.1109/ICDAR.1995.601971.

[42] PENG Q, PAN Y, WANG W, et al. ERNIE-layout: Layout knowledge enhanced pre-training for visually-rich document understanding[C/OL]//GOLDBERG Y, KOZAREVA Z, ZHANG Y. Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 3744-3756. https://aclanthology.org/2022.findings-emnlp.274/. DOI: 10.18653/v1/2022.findings-emnlp.274.

[43] LI Q, LI Z, CAI X, et al. Enhancing visually-rich document understanding via layout structure modeling[C/OL]//MM '23: Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2023: 4513–4523. https://doi.org/10.1145/3581783.3612327.

[44] JAUME G, EKENEL H K, THIRAN J P. Funsd: A dataset for form understanding in noisy scanned documents[A/OL]. 2019. arXiv: 1905.13538. https://arxiv.org/abs/1905.13538.

[45] XU Y, LV T, CUI L, et al. XFUND: A benchmark dataset for multilingual visually rich form understanding[C/OL]//MURESAN S, NAKOV P, VILLAVICENCIO A. Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, 2022: 3214-3224. https://aclanthology.org/2022.findings-acl.253/. DOI: 10.18653/v1/2022.findings-acl.253.

[46] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [C]//NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2020.

[47] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways[A/OL]. 2022. arXiv: 2204.02311. https://arxiv.org/abs/2204.02311.

[48] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[A/OL]. 2023. arXiv: 2302.13971. https://arxiv.org/abs/2302.13971.

[49] SUN Y, WANG S, LI Y, et al. Ernie: Enhanced representation through knowledge integration[A/OL]. 2019. arXiv: 1904.09223. https://arxiv.org/abs/1904.09223.

[50] SUN Y, WANG S, LI Y, et al. Ernie 2.0: A continual pre-training framework for language understanding[A/OL]. 2019. arXiv: 1907.12412. https://arxiv.org/abs/1907.12412.

[51] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[A/OL]. 2023. arXiv: 1910.10683. https://arxiv.org/abs/1910.10683.

[52] CHEN M, TWOREK J, JUN H, et al. Evaluating large language models trained on code[A/OL]. 2021. arXiv: 2107.03374. https://arxiv.org/abs/2107.03374.

[53] OPENAI, ACHIAM J, ADLER S, et al. Gpt-4 technical report[A/OL]. 2024. arXiv: 2303.08774. https://arxiv.org/abs/2303.08774.

[54] DRIESS D, XIA F, SAJJADI M S M, et al. Palm-e: An embodied multimodal language model[A/OL]. 2023. arXiv: 2303.03378. https://arxiv.org/abs/2303.03378.

[55] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[A/OL]. 2023. arXiv: 2307.09288. https://arxiv.org/abs/2307.09288.

[56] GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The llama 3 herd of models[A/OL]. 2024. arXiv: 2407.21783. https://arxiv.org/abs/2407.21783.

[57] TEAM G, GEORGIEV P, LEI V I, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[A/OL]. 2024. arXiv: 2403.05530. https://arxiv.org/abs/2403.05530.

[58] TEAM G, GEORGIEV P, LEI V I, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[A/OL]. 2024. arXiv: 2403.05530. https://arxiv.org/abs/2403.05530.

[59] GLM T, :, ZENG A, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools[A/OL]. 2024. arXiv: 2406.12793. https://arxiv.org/abs/2406.12793.

[60] QWEN, :, YANG A, et al. Qwen2.5 technical report[A/OL]. 2025. arXiv: 2412.15115. https://arxiv.org/abs/2412.15115.

[61] DEEPSEEK-AI, LIU A, FENG B, et al. Deepseek-v3 technical report[A/OL]. 2025. arXiv: 2412.19437. https://arxiv.org/abs/2412.19437.

[62] DEEPSEEK-AI, GUO D, YANG D, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[A/OL]. 2025. arXiv: 2501.12948. https://arxiv.org/abs/2501.12948.

[63] WANG P, BAI S, TAN S, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution[A/OL]. 2024. arXiv: 2409.12191. https://arxiv.org/abs/2409.12191.

[64] BAI J, BAI S, CHU Y, et al. Qwen technical report[A/OL]. 2023. arXiv: 2309.16609. https://arxiv.org/abs/2309.16609.

[65] YANG A, YANG B, HUI B, et al. Qwen2 technical report[A/OL]. 2024. arXiv: 2407.10671. https://arxiv.org/abs/2407.10671.

[66] LU P, BANSAL H, XIA T, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts[A/OL]. 2024. arXiv: 2310.02255. https://arxiv.org/abs/2310.02255.

[67] MATHEW M, KARATZAS D, JAWAHAR C V. Docvqa: A dataset for vqa on document images[A/OL]. 2021. arXiv: 2007.00398. https://arxiv.org/abs/2007.00398.

[68] TANG J, LIU Q, YE Y, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering[A/OL]. 2024. arXiv: 2405.11985. https://arxiv.org/abs/2405.11985.

[69] 陈梦泽. 人类视觉注意力的发展与分析[J/OL]. 吉林省教育学院学报 (上旬), 2012, 28(12): 139-140. DOI: 10.16083/j.cnki.1671-1580.2012.12.034.

[70] 王文冠, 沈建冰, 贾云得. 视觉注意力检测综述[J/OL]. 软件学报, 2019, 30(02): 416-439. DOI: 10.13328/j.cnki.jos.005636.

[71] SCHEITER K, GOG T. Using eye tracking in applied research to study and stimulate the processing of information from multi-representational sources[J/OL]. Applied Cognitive Psychology, 2009, 23: 1209 - 1214. DOI: 10.1002/acp.1524.

[72] HESSELS R S, HOOGE I T. Eye tracking in developmental cognitive neuroscience – the good, the bad and the ugly[J/OL]. Developmental Cognitive Neuroscience, 2019, 40: 100710. https://www.sciencedirect.com/science/article/pii/S187892931930297X. DOI: https://doi.org/10.1016/j.dcn.2019.100710.

[73] KILIAńSKA-PRZYBYłO G, GROTEK M. Eye-tracking: A guide for applied linguistics research[J/OL]. Folia Linguistica, 2021, 55(1): 275-279. https://doi.org/10.1515/flin-2020-2054. DOI: doi:10.1515/flin-2020-2054.

[74] 杨维春. 阅读认知理论指导下的法语阅读教学改革[J]. 中国法语专业教学研究, 2013(00): 243-249.

[75] 闫国利, 熊建萍, 臧传丽, 等. 阅读研究中的主要眼动指标评述[J]. 心理科学进展, 2013, 21(04): 589-605.

[76] PENG Q, PAN Y, WANG W, et al. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding[A/OL]. 2022. arXiv: 2210.06155. https://arxiv.org/abs/2210.06155.

[77] YE M, ZHANG J, ZHAO S, et al. Dptext-detr: Towards better scene text detection with dynamic points in transformer[A/OL]. 2022. arXiv: 2207.04491. https://arxiv.org/abs/2207.04491.

[78] WANG H, WANG Q, LI Y, et al. DocTrack: A visually-rich document dataset really aligned with human eye movement for machine reading[C/OL]//BOUAMOR H, PINO J, BALI K. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023: 5176-5189. https://aclanthology.org/2023.findings-emnlp.344/. DOI: 10.18653/v1/2023.findings-emnlp.344.

[79] 王晓峰杨亚东. 基于生态演化的通用智能系统结构模型研究[J/OL]. 自动化学报, 2020, 46(zdhxb-46-5-1017): 1017. http://www.aas.net.cn/article/doi/10.16383/j.aas.c170679.

[80] WOLFE J, CAVE K, FRANZEL S. Guided search: An alternative to the feature integration model for visual search[J/OL]. Journal of experimental psychology. Human perception and performance, 1989, 15: 419-33. DOI: 10.1037/0096-1523.15.3.419.

[81] LINDSAY G. Attention in psychology, neuroscience, and machine learning[J/OL]. Frontiers in computational neuroscience, 2020, 14: 29. DOI: 10.3389/fncom.2020.00029.

[82] PANG B, LI Y, LI J, et al. Tdaf: Top-down attention framework for vision tasks[A/OL]. 2020. arXiv: 2012.07248. https://arxiv.org/abs/2012.07248.

[83] KOCH K, MCLEAN J, SEGEV R, et al. How much the eye tells the brain[J/OL]. Current Biology, 2006, 16(14): 1428-1434. https://www.sciencedirect.com/science/article/pii/S0960982206016393. DOI: https://doi.org/10.1016/j.cub.2006.05.056.

[84] CARRASCO M. Visual attention: The past 25 years[J/OL]. Vision Research, 2011, 51(13): 1484-1525. https://www.sciencedirect.com/science/article/pii/S0042698911001544. DOI: https://doi.org/10.1016/j.visres.2011.04.012.

[85] BRADLEY M. Natural selective attention: Orienting and emotion[J/OL]. Psychophysiology, 2008, 46: 1-11. DOI: 10.1111/j.1469-8986.2008.00702.x.

[86] KATSUKI F. Bottom-up and top-down attention: Different processes and overlapping neural systems[J/OL]. The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry, 2013, 20. DOI: 10.1177/1073858413514136.

[87] BANERJEE S, GROVER S, SRIDHARAN D. Unraveling causal mechanisms of top-down and bottom-up visuospatial attention with non-invasive brain stimulation[J/OL]. Journal of the Indian Institute of Science, 2017, 97: 451 - 475. https://api.semanticscholar.org/CorpusID:186222083.

[88] YOHANANDAN S, SONG A, DYER A G, et al. Saliency preservation in low-resolution grayscale images[C]//FERRARI V, HEBERT M, SMINCHISESCU C, et al. Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 237-254.

[89] GASPELIN N, LUCK S J. “top-down” does not mean “voluntary” [J/OL]. Journal of Cognition, 2018. DOI: 10.5334/joc.28.

[90] GIESBRECHT B, WOLDORFF M, SONG A, et al. Neural mechanisms of top-down control during spatial and feature attention[J/OL]. NeuroImage, 2003, 19(3): 496-512. https://www.sciencedirect.com/science/article/pii/S1053811903001629. DOI: https://doi.org/10.1016/S1053-8119(03)00162-9.

[91] BALUCH F, ITTI L. Mechanisms of top-down attention[J/OL]. Trends in Neurosciences, 2011, 34(4): 210-224. https://www.sciencedirect.com/science/article/pii/S0166223611000191. DOI: https://doi.org/10.1016/j.tins.2011.02.003.

[92] XIA F, MCCORD M. Improving a statistical MT system with automatically learned rewrite patterns[C/OL]//COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland: COLING, 2004: 508-514. https://aclanthology.org/C04-1073/.

[93] COLLINS M, KOEHN P, KUcEROVá I. Clause restructuring for statistical machine translation[C/OL]//ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. USA: Association for Computational Linguistics, 2005: 531–540. https://doi.org/10.3115/1219840.1219906.

[94] NEUBIG G, WATANABE T, MORI S. Inducing a discriminative parser to optimize machine translation reordering[C/OL]//TSUJII J, HENDERSON J, PAŞCA M. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: Association for Computational Linguistics, 2012: 843-853. https://aclanthology.org/D12-1077/.

[95] LI C, GUO R, ZHOU J, et al. Pp-structurev2: A stronger document analysis system [J/OL]. CoRR, 2022, abs/2210.05391. https://doi.org/10.48550/arXiv.2210.05391.

[96] SEDGWICK P. Spearman's rank correlation coefficient[J/OL]. BMJ: British Medical Journal, 2014, 349: g7327. DOI: 10.1136/bmj.g7327.

[97] DEHLING H, VOGEL D, WENDLER M, et al. Testing for changes in kendall ' s tau [J/OL]. Econometric Theory, 2016, 33(6): 1352–1386. http://dx.doi.org/10.1017/S026646661600044X. DOI: 10.1017/s026646661600044x.

[98] MATHEW M, BAGAL V, TITO R P, et al. Infographicvqa[A/OL]. 2021. arXiv: 2104.12756. https://arxiv.org/abs/2104.12756.

[99] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of NeurIPS 2021. 2021.

[100] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. CoRR, 2017, abs/1706.03762.

[101] GUO M, XU T, LIU J, et al. Attention mechanisms in computer vision: A survey[J]. CoRR, 2021, abs/2111.07624.

[102] RAYNER K. Eye movements in reading and information processing: 20 years of research.[J]. Psychological bulletin, 1998, 124 3: 372-422.

[103] BAMMEL M, DE OLIVEIRA G S. Reading comprehension as embodied action: Exploratory findings on nonlinear eye movement dynamics and comprehension of scientific texts[C]//Annual Meeting of the Cognitive Science Society. 2023.

[104] DAMBACHER M, KLIEGL R, HOFMANN M, et al. Frequency and predictability effects on event-related potentials during reading[J]. Brain Research, 2006, 1084(1): 89-103.

[105] SOOD E, SHI L, BORTOLETTO M, et al. Improving neural saliency prediction with a cognitive model of human visual attention[C]//Annual Meeting of the Cognitive Science Society. 2023.

[106] CHEN S, KRUGER J L, DOHERTY S. Reading patterns and cognitive processing in an eye-tracking study of note-reading in consecutive interpreting[J/OL]. Interpreting, 2021, 23(1): 76-102. https://www.jbe-platform.com/content/journals/10.1075/intp.00050.che. DOI: https://doi.org/10.1075/intp.00050.che.

[107] SRIVASTAVA H. Zero shot crosslingual eye-tracking data prediction using multilingual transformer models[A/OL]. 2022. arXiv: 2203.16474. https://arxiv.org/abs/2203.16474.

[108] PAN Z, ZHANG G, PENG B, et al. Jnd-lic: Learned image compression via just noticeable difference for human visual perception[J/OL]. IEEE Transactions on Broadcasting, 2025, 71(1): 217-228. DOI: 10.1109/TBC.2024.3464413.

[109] BITEN A F, TITO R, MAFLA A, et al. Scene text visual question answering[C/OL]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 4290-4300. DOI: 10.1109/ICCV.2019.00439.

攻读硕士学位期间取得的研究成果

发表论文

[1] Hao Wang, **Qingxuan Wang**, Yue Li, Changqing Wang, Chenhui Chu, and Rui Wang. DocTrack: A Visually-Rich Document Dataset Really Aligned with Human Eye Movement for Machine Reading. Findings of the Association for Computational Linguistics: The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023). pages 5176–5189, Singapore. Association for Computational Linguistics.（第二作者，导师一作，CCF-B，第三章）

[2] **Qingxuan Wang**, Hao Wang, Huiran Zhang, Chenhui Chu, Rui Wang and Pinpin Zhu. Reading Between the Lines: How Eye-Tracking Data Can Inform Reading Strategies for Large Language Models. IEEE International Conference on Image Processing (ICIP 2025).（第二作者，CCF-C，第三章）

[3] **Qingxuan Wang**, Hao Wang, Huiran Zhang, Chenhui Chu, Rui Wang and Pinpin Zhu. Cognitive Insights into Document Comprehension: The Role of Reading Order and Visual Attention in Human and Large Language Models. Proceedings of the 47th Annual Meeting of the Cognitive Science Society (CogSci 2025).（第一作者，CCF-B，第四章）

致　谢

在论文完成之际，即将结束二十二年的求学之路，谨向所有给予我支持与帮助的师长、亲友致以最诚挚的谢意。

衷心感谢我的导师王教授与朱教授。从选题方向到论文定稿，两位老师以渊博的学识和严谨的治学态度为我指明方向，每一次讨论中的真知灼见都让我受益匪浅。特别感谢他们在疫情期间仍坚持线上指导，深夜回复的批注邮件和充满鼓励的语音留言，成为我学术道路上最坚实的支撑。

感谢课题组的同学以及室友。三年来，我们共同探讨技术难题，彼此支持鼓励，实验室里并肩作战的日夜、周末聚餐时的欢声笑语，都将成为最珍贵的回忆。特别感谢大家在数据收集阶段不辞辛苦的协助，让研究得以顺利完成。

感谢我的父母，你们以最朴素的爱筑起我追寻理想的基石。父亲沉默的守望与母亲温柔的叮咛，始终是漂泊求学路上最温暖的港湾。那些不曾言说的期待、深夜留灯的等待，以及电话两端心照不宣的牵挂，赋予了我直面挑战的勇气。你们教会我在泥土中扎根，在风雨里生长，这份坚韧与达观，早已成为刻在骨子里的精神基因。纸短情长，惟愿以不懈前行回报这份如山似海的恩情。

感谢上海大学计算机工程与科学学院提供的先进实验平台，同时感谢评审专家们提出的宝贵意见，让论文逻辑更加严密。

最后，感谢这个充满挑战与机遇的时代。学术之路虽道阻且长，但每一份善意都如同星光，照亮了探索的旅程。谨以此文，献给所有在平凡中孕育不凡的人们。

王庆旋

上海大学计算机工程与科学学院

2025 年 04 月 10 日