中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：22721570

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	人类阅读启发的强化学习阅读智能体研究

作　　者　王常青
学科专业　计算机应用技术
导　　师　王昊
完成日期　二〇二五年八月

姓　　名：王常青　　学号：22721570

论文题目：人类阅读启发的强化学习阅读智能体研究

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 08 月 13 日

姓　　名：王常青　　学号：22721570

论文题目：人类阅读启发的强化学习阅读智能体研究

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: 王常青

日期: 2025 年 08 月 13 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: 王常青　　导师签名:

日期: 2025 年 08 月 13 日　　日期: 2025 年 08 月 13 日

上海大学工学硕士学位论文

人类阅读启发的强化学习阅读智能体研究

作　　者: 王常青
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 8 月

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

Reinforcement Learning Reading Agents Inspired by Human Reading

Candidate: WangChangqing
Supervisor: WangHao
Major: Computer Application Technology

School of Computer Engineering and Science
Shanghai University
August, 2025

摘　要

在信息化时代的高速发展下，视觉富文档作为一种可以承载复杂信息的载体，在金融、法律、物流、学术等领域得到了大量的应用，这些应用包括 PDF，WORD，PNG 等多种形式，里面包含了丰富的视觉信息，布局信息和语义信息。文档的布局可以很大程度的反映文档想表达的含义，视觉元素和文本元素往往和布局信息一同组成了一个复杂信息的载体，但是随着视觉富文档的逐渐复杂化和视觉信息的增多，现有的文档理解模型难以对视觉富文档进行很好的理解，但是人类可以很好的利用视觉来对视觉富文档进行拆解分析，并且高效的获取自身需要的信息。为了提升文档理解模型的理解能力，参考人类进行文档布局感知，阅读顺序抽取就成为了迫在眉睫的任务。为了达成这一目标，需要克服语言、认知、技术等多方面的挑战，从而推动文档理解任务的发展。本文主要围绕于文档阅读顺序抽取任务在视觉富文档中的应用和下游任务中的推进进行了探讨，并且总结了以下的工作:

(1) 基于智能体自主探索的单页文档阅读顺序生成方法：现有的文档阅读顺序抽取方法主要依赖大量人工标注的数据，通常通过输入文档中各文本块的坐标信息、图像特征和文字内容，训练模型对每个待阅读内容进行顺序排序。然而，这类方法存在显著的局限性：首先，人工标注阅读顺序不仅耗时费力，而且标注过程枯燥繁琐，导致当前可用于训练的高质量标注数据极为稀缺。其次，传统的基于规则的方法通常依据固定的几何或布局规则来推断阅读顺序，缺乏对语义和上下文的理解能力，难以应对结构多样、格式灵活的真实文档，尤其在实际应用中表现出较差的泛化性和鲁棒性。尤其是在单页文档场景中，虽然视觉与布局信息更加集中，但复杂的排版、广告干扰、图文混排等仍对规则方法造成挑战。为解决上述问题，我们提出了一种基于强化学习的自主探索式文档阅读顺序抽取方法。具体而言，我们将文档的阅读行为建模为一个马尔可夫决策过程，设计了一种融合局部注意力与全局注意力机制的智能体结构，使其能够模拟人类在实际阅读过程中展现出的自适应策略与选择性注意行为。该强化学习智能体在接受下游任务反馈信号的指导下，逐步优化其阅读策略，实现对文档中关键内容的优先排序。实验结果表明，经该智能体处理后的阅读顺序能够显著提升下游文档理解任务（如问答、信息抽取等）的表现。同时，我

们在不同文档类型之间进行了迁移学习实验，结果表明该方法具有良好的迁移能力，验证了类似人类阅读策略的行为建模对于提升文档理解模型性能的积极作用。

(2) 基于人类对齐的多页文档阅读顺序生成方法：多页结构化文档中常出现跨页文本、表格截断等复杂布局，使传统 OCR 引擎基于线性顺序策略难以建模其空间与语义间的高度联动性，从而影响下游理解任务的效果。目前的阅读顺序抽取方法主要集中于单页文档，缺乏具有人类认知特征的多页阅读顺序标注数据，限制了强化学习等方法在多页文档场景下的性能提升。为此，我们构建了一个覆盖不同长度文档（单页、中短篇、超长篇），并由具备文档分析经验的标注者模拟真实眼动轨迹所标注的高质量阅读顺序数据集。该数据集有效刻画了人类自然阅读行为，为多页文档阅读顺序建模提供了坚实基础。在此基础上，我们结合多模态预训练模型的跨模态建模能力，利用人类标注数据进行微调，提出了一种具备结构感知与语义理解能力的阅读顺序抽取大模型。实验表明，该模型在多个下游任务中均显著优于不考虑阅读顺序的基线方法，验证了人类对齐策略在提升长文档理解效果方面的有效性。同时，跨文档迁移实验也展示了模型在不同领域和结构下的良好泛化能力，为多页文档的结构化理解提供了一种通用且高效的解决方案。

综上所述，本文围绕视觉富文档中阅读顺序建模的核心挑战，提出了基于强化学习的单页自主探索方法与基于人类认知对齐的多页建模方法，拓展了文档结构理解的研究路径，推动了文档阅读顺序建模从规则驱动向智能策略驱动的转变，为多模态文档智能理解提供了新范式与实践基础。

关键词:视觉富文档理解；强化学习；阅读顺序抽取；大模型智能体

ABSTRACT

With the rapid advancement of the information age, visually-rich documents (VRDs) have become widely used carriers of complex information in domains such as finance, law, logistics, and academia. These documents appear in various formats, including PDF, Word, and PNG, and contain rich visual, layout, and semantic information. Document layout often reflects the author's intended semantics, and visual and textual elements jointly convey intricate meanings through structural organization. However, as the complexity of VRDs increases and the amount of visual information grows, existing document understanding models struggle to effectively interpret such content. In contrast, humans can leverage visual perception to analyze VRDs and efficiently extract the needed information. To improve the capability of document understanding models, it is imperative to develop layout-aware models that mimic human reading behavior, making the task of reading order extraction an urgent problem. Addressing this challenge requires overcoming barriers in language, cognition, and technology to promote the development of document understanding.

This paper focuses on the modeling of reading order in VRDs and its impact on downstream tasks, presenting two key contributions:

(1) Agent-based autonomous reading order generation for single-page documents: Current reading order extraction methods mainly rely on large-scale human-annotated datasets. These methods typically train models to sort text blocks based on spatial coordinates, visual features, and textual content. However, they suffer from several limitations: manual annotation is time-consuming and tedious, leading to a scarcity of high-quality training data; rule-based methods often follow fixed geometric heuristics and lack semantic or contextual understanding, resulting in poor generalization to complex layouts. Even for single-page documents, where visual and layout information is relatively concentrated, challenges such as complex typesetting, visual clutter, and mixed content remain problematic. To address these issues, we propose a reinforcement learning-based autonomous exploration approach for reading order extraction. We formulate the reading process as a Markov Decision

Process and design an agent architecture that integrates local and global attention mechanisms to simulate human adaptive reading strategies. Guided by feedback from downstream tasks, the agent iteratively optimizes its reading strategy, prioritizing key content. Experimental results demonstrate that the generated reading order significantly enhances performance in document understanding tasks such as question answering and information extraction. Furthermore, transfer learning experiments across various document types confirm the method's robustness and the effectiveness of human-like strategy modeling.

(2) Human-aligned reading order generation for multi-page documents: Multi-page structured documents often feature cross-page text, table splits, and other complex layouts, making linear-order strategies from conventional OCR engines inadequate for capturing the spatial-semantic interplay, thus impairing downstream task performance. Existing methods largely focus on single-page documents and lack high-quality multi-page reading order data with human cognitive alignment, limiting the performance of reinforcement learning methods in such scenarios. To address this, we constructed a high-quality reading order dataset that covers documents of varying lengths (single-page, mid-length, and long-form), annotated by document analysis experts simulating natural eye movements and reading behavior. This dataset captures realistic human reading patterns and serves as a solid foundation for multi-page reading order modeling. Building upon this, we fine-tune a multimodal pretrained model using the human-aligned annotations and propose a large-scale model capable of capturing both structural and semantic signals. Experimental evaluations on downstream tasks demonstrate that the proposed model substantially outperforms baselines that ignore reading order, validating the efficacy of human-aligned strategies. Cross-domain transfer experiments further show strong generalization, highlighting the model' s potential as a universal solution for structured understanding of multi-page VRDs.

In summary, this work addresses the core challenges in reading order modeling for VRDs. We propose a reinforcement learning-based method for single-page autonomous reading strategy learning, and a human-aligned approach for multi-page reading order modeling. These contributions advance the field of document structural understanding and promote the transition from rule-based to strategy-driven reading order extraction, laying a new foundation for multimodal intelligent document understanding.

Keywords: Visually-rich document understanding; Reinforcement learning; Reading order extraction; Large model agent

目 录

第一章　绪论

1.1　研究背景和意义

随着人工智能技术的不断发展，视觉富文档（Visually-rich Documents，VRDs）[1]理解任务逐渐成为文档智能处理领域的重要研究方向。与传统纯文本文档不同，视觉富文档融合了文字内容、视觉排版、图像元素等多模态信息，其结构复杂、布局多变，广泛存在于发票、表格、宣传页、合同、学术论文等实际场景中。如何有效地解析这些信息，建立文档内容的语义关联，对于信息抽取、文档问答、文档分类等下游任务具有重要意义。

在现实世界中，VRDs 在各行各业中都具有广泛的应用场景。以票据识别为例，一张报销单可能包含打印文字、手写备注、红色印章、二维码、表格边框等多模态信息；又如法律文书中，段落层级结构和条文编号也构成了语义理解的重要基础。这些信息在视觉上呈现出明确的组织关系，人类在阅读时可以通过布局快速捕捉关键信息。但这也意味着，VRDs 的解析远比传统文本更为复杂。仅仅依赖纯文本建模往往难以捕捉其结构语义，造成信息丢失或理解偏差。并且 VRDs 的布局信息也远比传统的文本要复杂，存在着大量的图片模态信息作为辅助理解的补充。

随着人工智能特别是计算机视觉与自然语言处理技术的快速发展，传统依赖人工手动处理文档的方式逐渐暴露出其低效、高成本的问题。尽管人工标注方式具有较高的准确率，但在处理海量文档时面临着严重的时间与人力资源瓶颈，尤其在金融、政务、法律、物流等领域中，大量文档需要在短时间内完成解析和理解。因此，自动化的文档智能理解技术（Document Intelligence）应运而生，并迅速成为文档处理领域的重要研究方向。

目前，针对 VRDs 的文档理解任务已被广泛应用于多个关键领域，典型的应用包括：财务票据识别与报销单据解析、简历自动筛选与信息抽取、学术作业比对、法律政策文档的解析、合同审查、物流单据的跟踪与问答等。这些任务本质上涉及多模态信息的联合建模与推理，既需要对文档的图像形式进行精准感知，也需要对语言语义进行深层次理解。如图1.1所示，所以我们需要对多种多样的文档进行阅读顺序的解析 传统的规则引擎与模板方法很难适配通用场景，因此，近年来大语言模型（Large

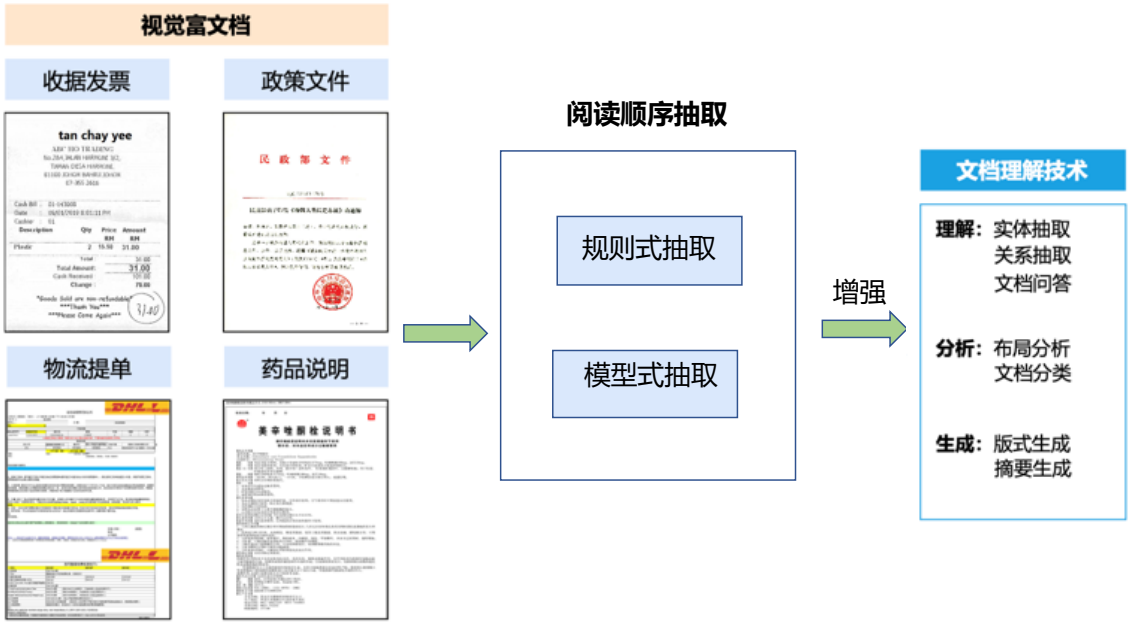


图 1.1 阅读顺序抽取的意义

Language Model，LLMs）逐渐成为主流解决方案。语言模型可以通过大规模预训练掌握语言结构与知识信息，借助光学字符识别（Optical Character Recognition，OCR）技术将图像中的文本转换为结构化序列后输入模型，执行命名实体识别（Named Entity Recognition，NER）[2]、关系抽取[3]、问答、分类等任务。但这种方法存在一个核心问题：OCR 引擎所提取的文本顺序往往不符合人类的阅读习惯。不同的数据集和文档类型具有不同的默认阅读顺序，而 OCR 引擎常常采用从左到右、从上到下的简单规则，无法还原表格内部、图文混排文档中真实的语义顺序。并且在面对多栏场景的文档时，采用默认的 OCR 引擎提取得到的阅读顺序往往会造成文本错位，严重干扰大模型进行理解过程。

这种不合理的阅读顺序对语言模型的理解能力构成了直接干扰，尤其是在涉及 key-value 对（即字段名与其值）、标题-正文结构等强依赖上下文的任务中更为明显。例如，在财务报表中，“金额”这一字段应当与紧随其后的数值对应，但 OCR 输出可能将字段与其值打乱或穿插其它内容，导致模型理解偏差。为了解决这一问题，越来越多的研究开始探索如何优化语言模型的输入顺序，使其更贴合人类的认知逻辑。

随着大模型推理与生成能力的不断增强，利用大语言模型对文档内容进行端到端理解成为可能。我们可以通过调整阅读顺序、引入多模态输入（如布局信息、视觉特征等）、优化提示 (prompt) 设计等方式，引导模型从更加合理的角度解析文档。特别是在视觉富文档中，合理构建输入顺序不仅能显著提升信息提取的准确率，还能

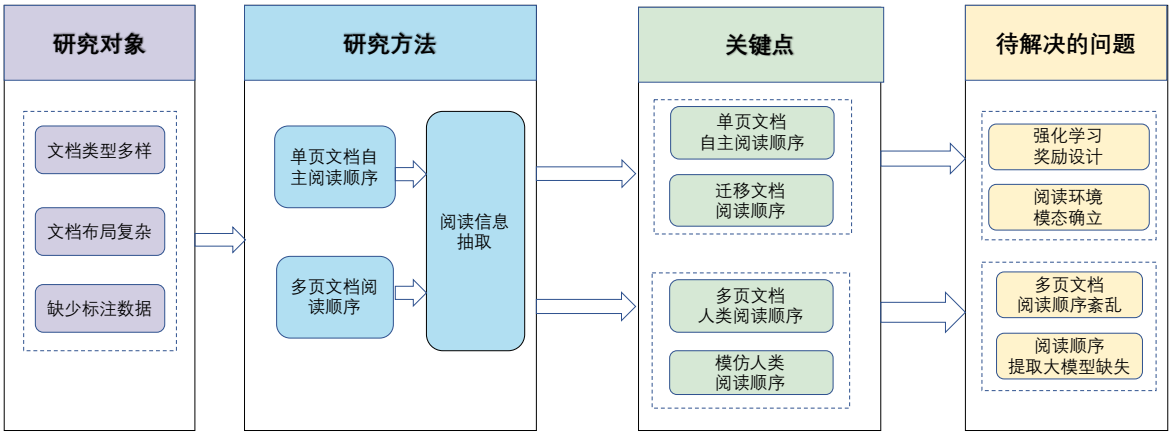


图 1.2 本文的研究内容

挖掘出更深层次的语义关联关系。

因此，如何合理的提取出文档的阅读顺序，充分增强下游任务的理解能力，让大模型可以根据阅读顺序进行相关任务的解析和完成，成为了当前文档智能领域的重要研究方向之一。

1.2 研究问题

针对上述内容，本研究主要关注于输入到模型中的阅读顺序的问题。通过调整输入到模型中的阅读顺序，尽可能在预训练完成的模型中，最大化提升模型的理解能力。并且探究适合大模型理解的阅读顺序和人类阅读顺序之间的差异性，所以本文主要聚焦于以下两点

(1) 如何利用智能体自主探究模型的阅读顺序从而增强下游语言模型能力

研究发现，由于文档的多样性和复杂性，识别文档是一个费时费力的工作，单纯的文本信息无法应对现在复杂的文档理解任务，我们需要一种多模态的方法，通过联合建模的方式来考虑多模式的一致性，包括文本、视觉和布局。现有的多模态信息抽取模型例如 LayoutLM[4-6] 系列模型和 StructTexT[7-9] 系列模型等。这些模型可以从多模态中可以得到细粒度的表示，但是这些模型缺乏从给定文档中产生合理阅读顺序的能力。

因此，他们往往采用的是简单的阅读顺序，例如简单的 OCR 引擎中的先左右后上下的规则式解析过程，但是这些阅读顺序往往不能满足复杂多变的文档布局的需求，并且在文档中进行大量的阅读标注是成本比较高的，强化学习的方法也可以在这种排序场景中获得较好的利用，例如 PointerNetwork[10-11] 系列，经

过强化学习的微调可以很好的进行排序，但是缺少对阅读场景的环境建模，以及奖励函数，针对上述问题，如何构建强化学习智能体动态阅读单页的文档阅读顺序就成为了我们现在需要解决的问题。

(2) 如何建立多页文档的阅读顺序抽取模型

随着大模型技术的发展，对于多页文档的阅读顺序解析成为了新的需求，但是业界现有的阅读顺序解析模型，例如 LayoutReader[12],DocTrack[13] 等模型大多只能解析单页的文档阅读顺序，并且业界存在的数据集例如 DocTrack 数据集和 ReadingBank 数据集等大多集中于单页的人类阅读顺序，那么针对于多页文档阅读中存在的跨页信息损失，表格解析失败等问题迫切的需要被解决，并且如何构建多页人类阅读顺序和如何抽取多页的阅读顺序抽取智能体没有相关的研究和框架。

1.3 研究内容

基于上述讨论，本文将会从两个方面进行文档理解任务的阅读顺序研究并尝试解决自主探索语言模型阅读顺序、如何构建多页文档阅读顺序数据集和通用阅读顺序抽取大模型。如图1.2所示

1. 为了探究语言模型自主的阅读顺序，我们利用语言模型在下游的表现作为强化学习的奖励，并且充分利用人类阅读的过程，将阅读过程需要使用的文字，图片，位置信息作为强化学习智能体[14]的多模态输入，通过强化学习强大的特征提取融合能力，将三种模态充分融合，让语言模型可以自适应的调整自身需要的阅读顺序，对语言模型的理解能力进行增强。
2. 为了得到多页的阅读顺序标注数据集，我们征集了十几位研究生志愿者，对超长文档、中短文档和单页文档进行了阅读顺序的采集，并且进行了数据的清洗和处理。为了充分利用该标注阅读顺序，我们进行了多模态大模型的 GRPO[15] 训练，将人类的阅读顺序作为对齐的目标，设计了奖励函数微调语言模型作为多页文档阅读顺序抽取模型。

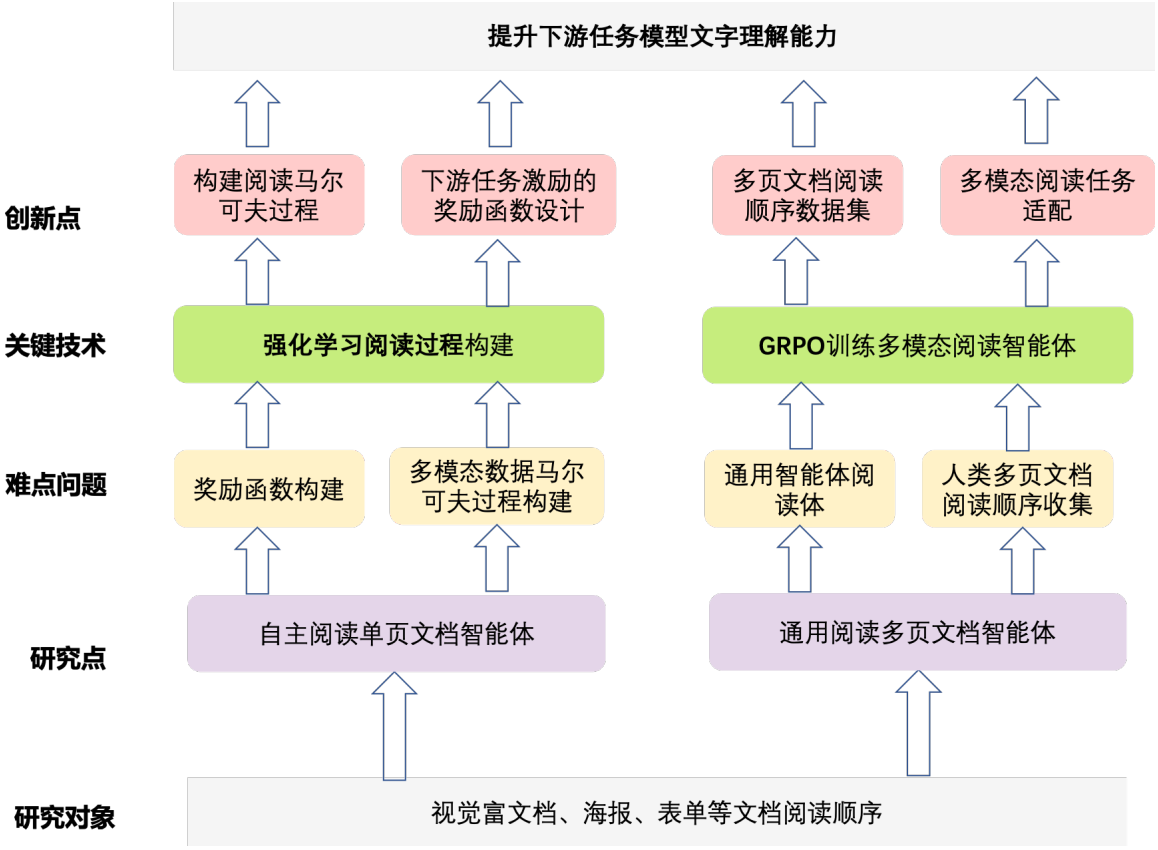


图 1.3　本文的主要创新点

1.4　创新点

本文针对于多模态文档的阅读顺序抽取任务，从多模态强化学习阅读智能体构建，通用阅读智能体训练两方面进行研究。图1.3展示了本文的创新点

(1) 通过下游任务为奖励为导向，模拟人类阅读过程构建强化学习阅读智能体阅读过程：现有的阅读顺序抽取方式大多依赖于人类标注数据，基于类似 Transformer[16]的架构，例如 LayoutReader 模型，对 PDF 文档和 XML 文档进行端到端的标签预测，DocTrack 数据集中提出了一种基于先后排序的 token 比较方法，即让 LayoutLM 系列模型作为基础模型，通过预测两个 Token，谁在前谁在后，对所有的 Token 进行了标注，最后再使用类似冒泡排序的方式对所有的 token 进行排序，从而得到对应的阅读顺序。或者类似图网络的方式，复旦大学提出的 TPP[17]的方式也有效的利用了人类标注的数据，将文档阅读顺序问题看作是图中节点的层次预测问题。但是这些阅读顺序提取网络都有着相似的问题，例如这些网络模型都强依赖于人类标注的数据，如果文档的结构发生巨大变化，例如增加了图片的箭头引导阅读过程时，这些阅读顺序提取模型在这些任务中就会表现的非常有限。为了解决这个问题，

我们引入强化学习的方式，让智能体自己进行阅读过程，从而脱离人类标注的约束，可以通过和文档的交互从而增强智能体的阅读理解过程。第二这些阅读顺序的感知过程往往没有充分利用图片的多模态的理解能力，我们的方法通过充分融合各种模态来对阅读过程进行建模，从而让智能体可以类似于人类，进行局部阅读和全局阅读相结合，从而让阅读智能体可以充分增强理解阅读内容。

(2) 多页文档阅读顺序数据集与通用阅读顺序抽取大模型的构建：随着大语言模型技术的快速发展，模型在多种自然语言理解任务中展现出强大的通用能力。当前的多模态或结构感知大模型仅依赖文本内容及其空间坐标信息，便能在实体识别、关系抽取、文档分类、信息匹配、问答等任务中取得令人瞩目的表现。然而，在复杂文档结构中，模型的理解效果往往严重依赖于输入信息的结构化程度，其中**是否具备合理的阅读顺序**成为影响文档理解性能的关键因素之一。尽管大模型拥有强大的推理与语言建模能力，若其接收到的内容顺序不符合人类的真实阅读逻辑，往往会造成上下文割裂、语义混乱等问题，进而影响下游任务的表现。

在当前的多页文档理解场景中，传统 OCR 引擎通常采用固定的**自上而下、从左到右**的线性规则对文档进行顺序提取，难以适应实际文档中常见的表格跨页、段落跳转、分页连接等复杂结构。此外，基于模型自动学习的阅读顺序策略在缺乏高质量监督信号的情况下，也常常会陷入歧义解析、信息遗漏等问题，严重制约了多页文档场景下文档理解系统的性能。因此，构建一个真实、合理的多页文档阅读顺序标注数据集，成为提升模型对复杂结构文档建模能力的关键基础。

为此，本文首次构建了一个覆盖多种真实文档结构、具有人类阅读行为标注的多页文档阅读顺序数据集。我们招募了多名具备文档分析经验的研究生志愿者，对超过 500 份实际多页文档进行了细致的人工标注，标注内容涵盖了段落顺序、跨页跳转、视觉引导路径等多种人类阅读行为，并辅以眼动轨迹记录以增强标注的自然性与一致性。在此基础上，我们进一步将该数据集用于指导多模态大模型进行 GRPO 微调训练。具体而言，我们以人类眼动数据为行为参考信号，设计奖励函数引导模型学习接近人类阅读策略的内容排序路径，从而实现多页文档阅读顺序的学习与泛化。

实验结果表明，所提出的通用阅读顺序抽取大模型不仅在多个文档理解任务（如长文档问答、信息抽取）中显著提升性能，而且在不同类型和结构的文档之间表现出良好的迁移能力。这一成果充分验证了：引入高质量人类阅读顺序数据监督、结

合强化学习方法对多模态大模型进行微调，是实现复杂文档结构下通用文档理解模型构建的重要路径。该工作为后续多页文档场景下的通用智能文档系统发展奠定了坚实的基础。

1.5 本文的架构

本文围绕多模态文档中的阅读顺序抽取问题展开了系统性研究，旨在解决当前模型对复杂文档结构适应性差、对大量人工标注依赖严重等现实难题。首先，本文系统介绍了多模态文档阅读顺序抽取任务的研究价值与核心挑战，指出传统方法在处理结构复杂、多页交错、视觉信息丰富的文档时，难以准确建模人类真实的阅读行为。此外，现有方法多依赖静态的人类标注数据，限制了其在实际应用场景中的泛化能力与灵活性。

针对上述问题，本文提出了一种基于强化学习的阅读顺序抽取方法，将文档阅读建模为一个马尔可夫决策[18]过程，利用下游任务（如问答、信息抽取）的性能表现作为强化学习智能体的奖励信号，引导其自主学习合理的阅读路径。该方法不仅能够跳脱出对人类固定顺序的依赖，还能够挖掘出不同于人工经验或规则式设计的新型阅读路径，从而为文档理解任务提供更加高效的结构输入。

为了进一步增强阅读顺序抽取模型的通用性，本文在强化学习框架基础上，引入了当前主流多模态大模型，通过引入 GRPO 机制，将人类阅读顺序作为参考信号，引导大模型进行行为对齐与策略优化。实验结果表明，经过微调的大模型在多个文档理解下游任务中均取得了优越性能，显示出良好的泛化能力与任务适应性。

此外，为了解决多页文档中常见的跨页跳转、信息割裂等问题，本文自主构建了一个高质量的多页文档阅读顺序标注数据集。我们组织了十余位具备文档理解经验的研究生志愿者，对 500 余份真实文档进行了阅读顺序标注，采集了贴近人类真实阅读路径的行为数据，并在此基础上训练了具备跨页感知能力的通用阅读顺序抽取智能体，在多个多页文档场景下取得了优异表现。

本文的整体章节安排如下：

- **第一章绪论**：介绍了阅读顺序抽取任务的重要性与研究背景，系统梳理了当前该领域面临的两个核心挑战，并提出了本文的研究目标与创新点。
- **第二章相关理论和研究方法**：综述了现有阅读顺序抽取方法的技术演进，并详

细介绍了规则驱动方法与数据驱动方法两大主流范式。同时，本章还引入了强化学习的基本原理，为后续方法设计提供理论支撑。

- **第三章基于智能体自主探索的单页文档阅读顺序生成方法**：提出了一种基于强化学习的自主阅读顺序抽取框架，通过将下游任务表现作为反馈信号，引导智能体优化阅读策略，实现在不同文档结构中的自适应排序。
- **第四章基于智能体人类对齐的多页文档阅读顺序生成方法**：围绕多页文档中的结构复杂性问题，构建了一个具备人类眼动对齐标注的阅读顺序数据集，并结合 GRPO 方法对多模态大模型进行微调，获得了适用于复杂多页文档的阅读顺序大模型。
- **第五章总结与展望**：总结了本文的主要工作与实验成果，分析了当前研究的不足，并对未来在多模态阅读顺序建模、跨任务迁移学习等方向提出了展望。

第二章　相关理论和研究方法

2.1　视觉富文档理解任务的研究现状

2.1.1　文字识别方法

光学字符识别是一项通过图像处理和光学扫描技术，将图像中的印刷体或手写文字内容转换为可编辑、可搜索的数字化文本的关键技术。其核心目标在于将文字图像转化为计算机可识别的字符，实现从视觉信息到结构化数据的转换。OCR 技术已在多个实际场景中广泛应用，例如：文档数字化（将纸质文档转为电子文本以便于存储和检索）、金融行业（票据识别、身份验证）、教育领域（试卷识别、作业批改）、物流领域（快递单号识别、货物信息录入）以及医疗场景（病历电子化、医学报告解析）等。

根据技术发展阶段，OCR 方法主要可分为三类：基于模板匹配的方法、基于统计学习的方法以及基于深度学习的方法。

(1) 基于模板匹配的 OCR 方法

早期的 OCR 方法主要依赖模板匹配技术，其基本思想是预先构建一个包含所有待识别字符的模板库，并将输入图像中的字符区域与模板逐一进行相似度比对[19]，最终选出最相似的模板所代表的字符作为识别结果[20]。这类方法具有实现简单、识别精度较高、计算效率快等优点。然而，其缺点也较为明显：一方面，模板库的构建成本高，需穷举所有字符样式；另一方面，方法的泛化能力较差，对于未出现于模板中的字符或结构复杂、形变显著的手写体字符难以准确识别。因此，该方法难以适应实际应用中多样化和复杂化的文本场景。

尽管如此，模板匹配法为后续文字识别技术的发展奠定了基础。随着文档结构复杂性的提升，其局限性日益显现，促使研究者转向更具鲁棒性和泛化能力的统计学习方法。

(2) 基于统计学习的 OCR 方法

伴随机器学习的发展，OCR 技术逐渐转向依赖统计模型进行字符识别。该类方法以人为设定的字符形态特征为基础，通过提取字符的结构、几何或统计特征，并利用分类模型（如支持向量机 SVM、隐马尔可夫模型 HMM）对字符进行识别。主

要流程包括：

- **结构特征提取：**利用字符边缘、轮廓或频域信息，如边界特征法或傅里叶变换特征[21-22]；
- **几何分布特征提取：**如二维直方图投影法、区域网格统计法等，从字符的空间分布中提取规律；
- **统计特征提取：**如灰度直方图和纹理特征[23]，用于刻画字符图像的灰度和纹理属性。

完成特征提取后，可通过统计模型进行字符分类。这种方法在一定程度上提高了识别精度，并可利用上下文信息进行纠错。然而，其对手工设计特征高度依赖，难以全面表示字符的多维属性，限制了其在复杂场景下的性能。这一不足推动了深度学习方法在 OCR 领域的广泛应用。

(3) 基于深度学习的 OCR 方法

近年来，深度学习技术的迅猛发展为 OCR 带来了突破性进展。深度学习模型可自动从数据中学习特征，无需人工设计，极大地提升了字符识别的准确性与适应性。目前，主流的深度学习 OCR 方法主要包括基于卷积神经网络（CNN）、循环神经网络（RNN）以及注意力机制的模型。

- **卷积神经网络（CNN）**[24]：CNN 是当前 OCR 中最核心的特征提取模块，通过多层卷积、池化与非线性激活函数，CNN 能够高效提取图像中的局部与全局特征，广泛应用于文字检测和识别两个阶段。例如在古籍识别中，CNN 能有效捕捉字符的笔画结构与纹理信息，从而提升识别精度[25]。
- **循环神经网络（RNN）及其变体（如 LSTM 和 GRU）**[26]：RNN 能够处理序列数据，尤其适用于文字行内的字符序列识别。其上下文感知能力使得识别模型能更好地理解字符间的语义与结构关系，从而提升整体识别效果[27-28]。

2.1.2 基于预训练的文档理解模型

随着自然语言处理技术的发展，文档理解任务取得了巨大的进步，例如 BERT[29]模型通过掩码掉文本部分内容，让模型通过类似于完形填空的方式对掩码内容进行预测，从而学到上下游语义信息，GPT[30]相关模型则是采用生成式方式对文本进行生成，预测下一个词语的分布，从而提高自身的理解能力，这些预训练模型可以在大量的语料中进行模型的基本表示学习，并且在下游任务中进行微调就可

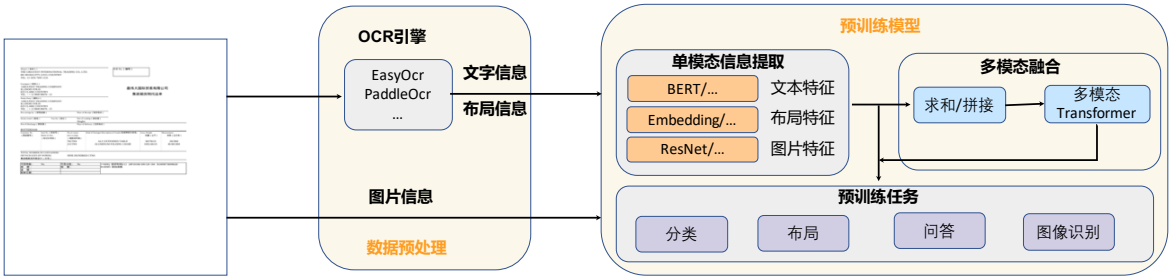


图 2.1 基于 Transformer 的多模态模型预训练架构

以很好的适配下游的任务，整体流程如图2.1所示，主要包括模态表示提取，多模态融合和下游任务微调三个部分。三个模块具体的含义是

- **模态表示提取：**指的是使用多种预训练模型对文档的不同模态进行编码
- **多模态融合：**指的是对于提取出的多种模态信息进行相加或者拼接进行融合。
- **下游任务微调:** 指的是在具体的下游任务中使用具体的任务标签对多模态融合的之后用于分类的模型进行参数的轻微调整。

BERT 模型：BERT 模型通过使用 WordPiece 分词算法构建了一个包含约 30,000 个子词单元的词汇表，并在此基础上学习文本的深层语义表示。每个输入序列的起始标记为 [CLS]，句子间使用 [SEP] 进行分隔，并使用 [PAD] 进行填充，以保证输入序列在长度上的一致性。BERT 的输入表示由三部分嵌入向量相加组成，分别为：Token Embedding（文本编码）、Position Embedding（一维位置编码）以及 Segment Embedding（片段编码），从而构建出完整的文本表示。

然而，在文档理解任务中，文字的空间位置信息往往蕴含着重要的结构语义。例如，在表单类文档中，键值对（key-value pair）通常以左右或上下排列的方式呈现，并具有固定的语义对应关系。对于视觉富文档，除了文本本身的语义外，其排版布局、字体样式、对齐方式等视觉特征同样承载着关键的语义线索，这些信息对于提升下游任务的性能具有重要作用。

因此，虽然 BERT 在文本语义建模方面具有强大的能力，但其缺乏对文本空间结构和视觉信息的建模能力，限制了其在文档图像场景下的应用效果。

LayoutLM 模型：LayoutLM 模型是微软提出的文档理解模型，为了解决 BERT 对于二维空间概念理解性差的问题，LayoutLM 引入了二维坐标编码信息和图片信息，融合了文本编码信息进行，同时学习文本、坐标、图片编码。基本架构如图2.2所示，可以从架构图中看出，LayoutLM 的文本信息由 BERT 提取得到，同时引入 2D 编码，分别为横坐标和纵坐标，并且用 Fast R-CNN[31]模型提取出图片模态的表示，将

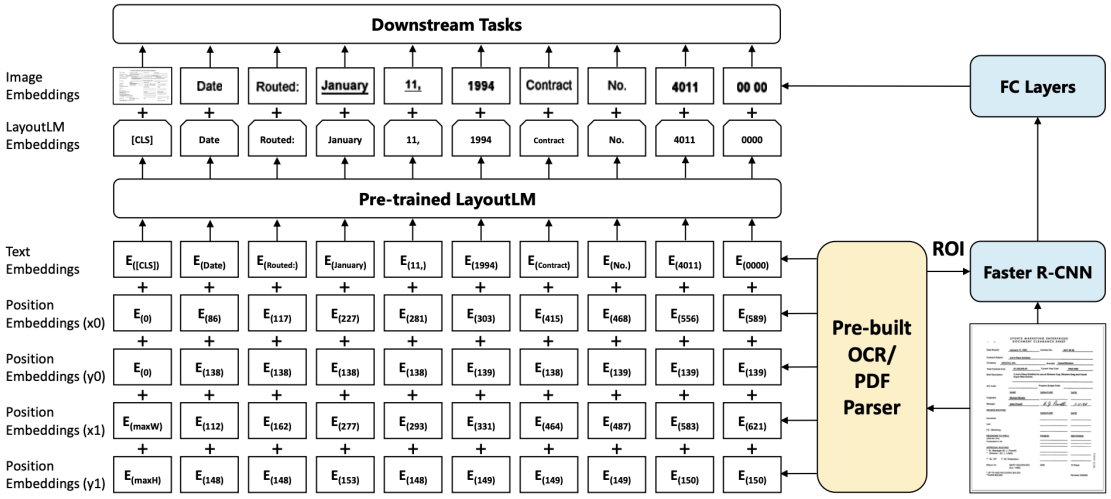


图 2.2　LayoutLM 模型架构[4]

三种模态进行融合。通过两个预训练任务进行基础参数的训练，分别是掩码视觉语言建模任务和多标签文档分类任务。在微调时引入了图片模态的信息，但是在下游任务中没有显著的性能提升，这是因为图片模态的提取能力一般，在多模态融合时能力有限，所以对性能的提升较为一般。

LayoutLMv2 模型： LayoutLMv2 是在 LayoutLM 的基础上进行改进的一种多模态预训练模型，引入了图像模态信息，使模型具备更强的视觉感知能力。该模型通过设计空间感知的自注意力机制，有效挖掘文档中二维布局的结构线索。同时，LayoutLMv2 将原有的多标签文档分类任务替换为文本—图像对齐任务，从而更充分地学习文本与图像之间的内在关联与交互关系。

在输入编码方面，LayoutLMv2 引入了模态类型编码，以区分来自文本和图像的不同模态信息，并为每种模态分别加入了一维的全局位置编码。模型首先分别对文本的语义表示、图像的视觉特征表示及其对应的布局表示进行加和，然后将两种模态的融合表示拼接作为 Transformer Encoder 的输入。其空间感知的自注意力机制将一维和二维位置偏置信息融入注意力计算过程中，从而提升对文档空间结构的建模能力。

与前代模型 LayoutLM 相比，LayoutLMv2 在多个任务中表现出更优的性能，尤其是在对视觉信息依赖较强的任务中，如文本与图像匹配、结构化信息抽取等。同时，LayoutLMv2 具备良好的通用性和扩展性，能够适应多种类型和格式的文档，具有较高的实际应用价值。

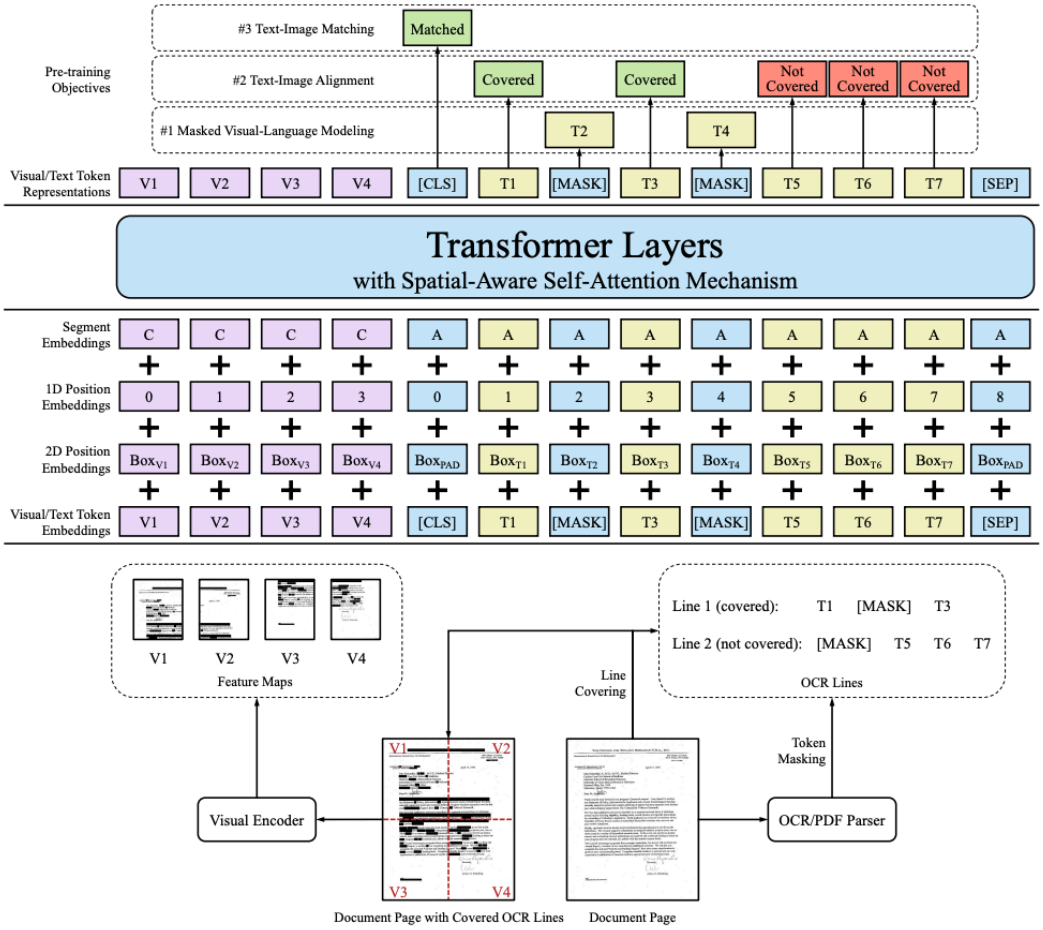


图 2.3　LayoutLMv2 模型架构[5]

2.1.3　基于大模型的文档理解方法

自 2017 年 Transformer 架构提出以来，大语言模型（LLMs）迅速发展。其引入的自注意力机制革新了自然语言处理方法，奠定了大模型演进的基础。2018 年，BERT（3 亿参数）与 GPT（1.17 亿参数）相继问世，分别代表了双向与单向建模的两种范式，初步展现出预训练模型的强大能力。

2019 年，模型规模持续扩张，OpenAI 发布 15 亿参数的 GPT-2[32]，支持多任务语言建模；百度同期推出 ERNIE[33-34] 系列，通过知识增强提升语义理解能力。2020 年，GPT-3[35] 实现 In-Context Learning，显著提升了模型的泛化推理能力；Google 提出 T5[36] 模型，统一 NLP 任务的建模方式。2021 至 2022 年，模型能力向多任务与代码生成延伸：OpenAI 发布 Codex[37]，Google 推出参数高达 5400 亿的 PaLM[38] 模型，Meta 则开源 650 亿参数的 LLaMA[39] 模型，极大推动了开源社区的发展。2023

年，LLMs 进入多模态融合阶段。OpenAI 推出支持文本与图像输入的 GPT-4（据称参数达 100 万亿）[40]，Google 发布多模态模型 PaLM-E[41]，百度升级文心一言（ERNIE 3.0），Meta 发布 LLaMA2（690 亿参数）[39]并于 7 月开源 LLaMA3（706 亿参数）[42]，首次在多个基准测试中超越闭源模型 Gemini 与 Claude，标志开源生态体系的快速崛起。2024 年，大语言模型的发展以推理能力提升与架构创新为核心。OpenAI 于 9 月发布 GPT-4.5，强化复杂逻辑处理能力；Google 基于混合专家（MoE）架构，相继推出 Gemini 1.5[43]（2024 年 2 月）与多模态 Gemini 2.0（2024 年 12 月），提升跨模态理解能力。与此同时，DeepSeek 系列快速迭代，2024 年 5 月发布多模态模型 DeepSeek-V3[44]，12 月推出具备高阶推理能力的 DeepSeek-R1[45]，尝试模拟人类思维过程以提升逻辑推理深度。进入 2025 年，OpenAI 相继发布 GPT-4o（2024 年 5 月）与 GPT-o3（2025 年 1 月），进一步加强模型在多模态融合与人格化交互方面的能力，推动类人智能的发展。

综上所述，LLMs 自单向序列建模起步，历经模型规模扩展、多任务与代码生成、多模态融合等关键阶段，逐步迈向具备推理能力、交互性与认知性的通用人工智能系统，并且持续拓展自然语言处理与跨模态理解的研究边界。

Algorithm 1 基于预排序模型的排序算法

Input: 输入 $\mathbf{B}$：原始 OCR 顺序的 bbox 级别的多模态特征序列 $[b_1, \ldots, b_n]$

Output: 输出：排序后的输出序列 $\mathbf{B}_r$

$\mathbf{B} \leftarrow [b_1, \ldots, b_n]$
for $i \leftarrow 0$ **to** n **do**
  for $j \leftarrow 0$ **to** $n - i - 1$ **do**
    $p \leftarrow$ 模型预测$(r_j : r_{j+1})$; // 调用预排序模型以确定优先顺序
    if $p < \theta$; // 阈值设置为 0.5
    then
      交换 r_j 和 r_{j+1}; // 与冒泡排序一致
    end
  end
end
return $\mathbf{B}_r$; // 新的排序序列

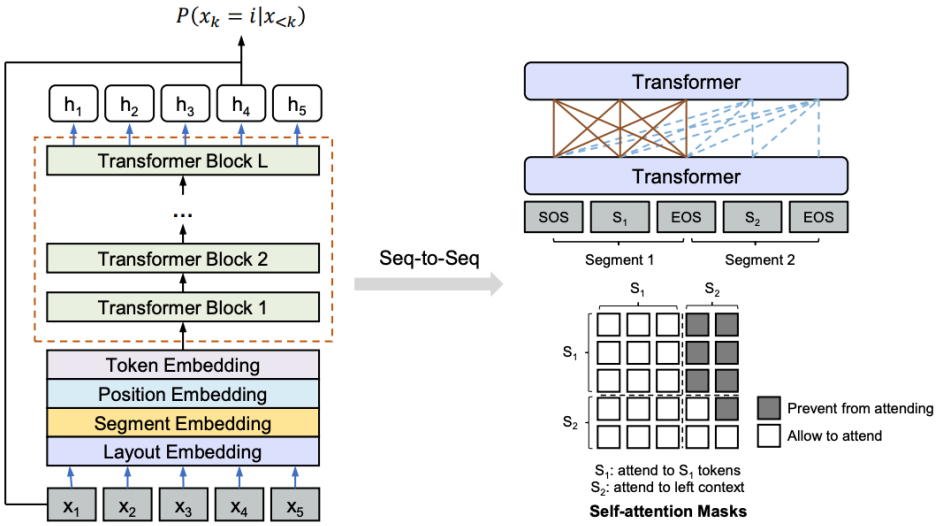


图 2.4　LayoutReader 架构图[12]

2.2　阅读顺序生成

2.2.1　基于规则式生成阅读顺序

基于规则式进行阅读顺序生成的方法可以简单分为几种，包括按照 Z 字顺序进行阅读顺序生成的方法这种方法获得的阅读顺序遵循先左右后上下的阅读规则[46-49]。算法流程如算法 1，张等人结合了排序算法，对文档阅读顺序采用了一种启发式的规则排序算法 XYCUT[50]这种算法优先考虑局部信息，并不是严格的按照左右上下的阅读顺序。

2.2.2　基于预训练模型的阅读顺序生成

王等人[12]首次提出利用微软 Word 文档的 XML 元数据，自动构建了包含 50 万张真实文档图像及其行阅读顺序标注的 ReadingBank 数据集，为大规模深度模型训练提供了基础。基于此数据集，作者设计了 LayoutReader 一个双流（文本与版式布局）输入的 Seq2Seq 模型，直接预测文本行或版块的阅读顺序，显著优于此前的规则和传统机器学习方法。在多种 OCR 引擎后处理实验中，该模型在行排序准确率上接近“人类标注”水平，提升了 OCR 输出的一致性与可读性，后续研究多次复现并拓展了 LayoutReader 在表单、票据和多列布局文档上的泛化能力，显示了大规模预训练与多模态融合在阅读顺序任务中的潜力。王等人 在 Findings of EMNLP 2023 中提出 DocTrack 数据集，通过眼动仪采集真实用户在阅读各类视觉文档（如表单、海报、发票）时的注视轨迹，并据此对文档中的文本区块进行先后关系标注，从而真正对齐人

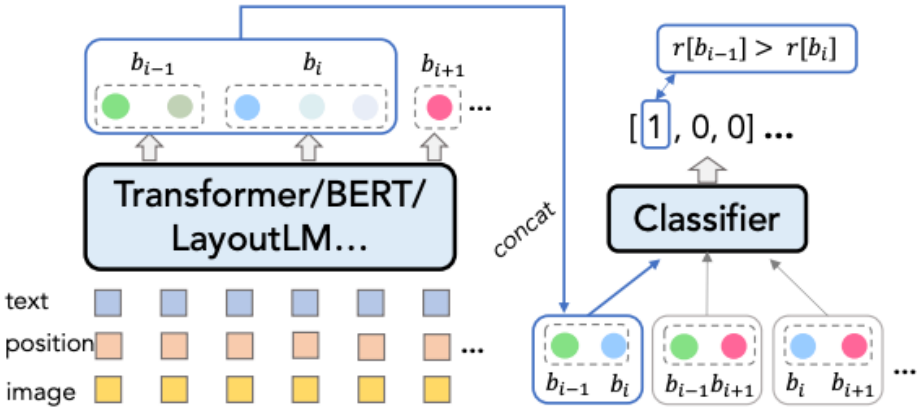


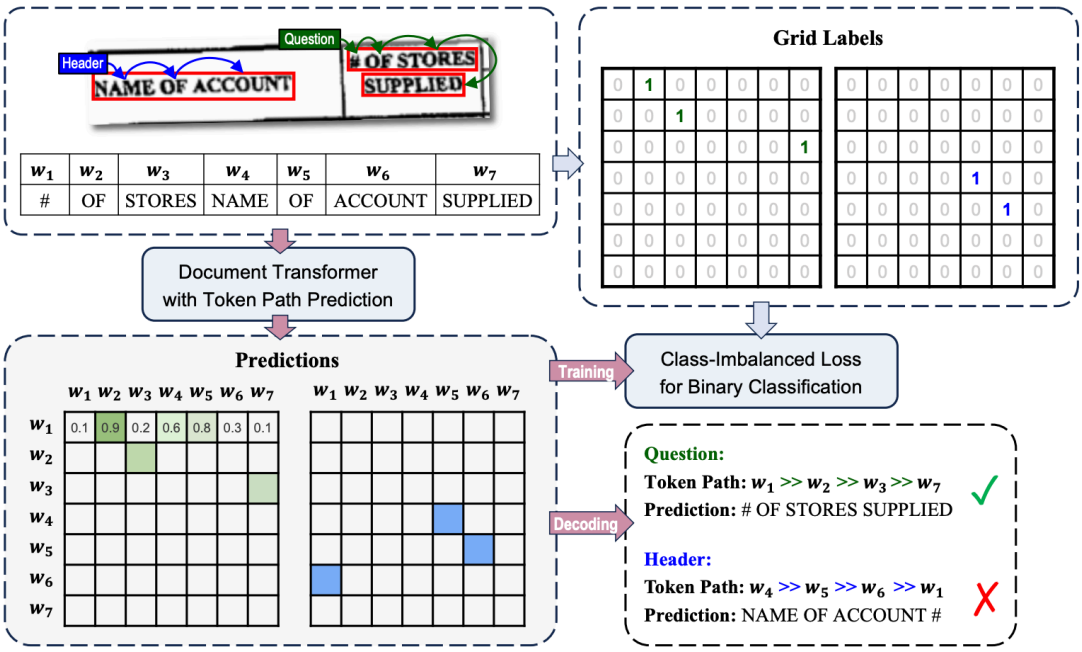
图 2.5 DocTrack 整体架构[13]

类的阅读习惯 DocTrack 包含 539 张图像，来源于 FUNSD[51]、SEABILL、InfoVQA[52] 等公开数据集，其预排序（preordering）流水线以眼动对（precedence pairs）为基础，利用图模型或 Transformer 重建全局阅读顺序，在信息抽取、表格理解等下游任务上均获得了显著提升该工作首次揭示了人类真实阅读顺序与现有 Document AI 模型预测顺序之间的差距，为未来结合认知信号的模型设计提供了重要参照。并且使用了类似于冒泡排序的方式，将文本之间两两比对阅读顺序的大小获得了文档的阅读顺序。

2.2.3 基于图神经网络的阅读顺序抽取方法

张等人在 EMNLP 2023 论文中，将阅读顺序抽取视为文档中任意两个 Token 间路径的预测问题，引入了 Token Path Prediction（TPP）模块：将文档视为完全有向图，模型需预测路径序列以重构全局顺序，TPP 方法不仅在阅读顺序检测上达到了最新性能，还与命名实体识别、实体链接（Entity Linking）等下游信息抽取任务进行了多任务联合训练，并提出了新的**顺序无关**评估指标，进一步验证准确的阅读顺序对实体识别质量的关键作用。该工作强调了阅读顺序作为文档多模态理解组件的核心价值，为未来基于图结构的联合建模提供了新思路。

李等人充分利用图神经网络的特性提出了 GraphLayoutLM[53]，并且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。该模型引入图顺序优化与图掩码优化两种策略，相较于传统方法，基于图结构显式构建文本节点间的层级关系与空间邻接关系，并通过图重排序算法动态调整



w_1	w_2	w_3	w_4	w_5	w_6	w_7
#	OF	STORES	NAME	OF	ACCOUNT	SUPPLIED

图 2.6　TPP 整体架构[17]

阅读顺序，结合图掩码机制优化自注意力，从而更精准地建模复杂文档中的逻辑结构。

2.3　RAG 系统相关研究

在当前的大语言模型应用中，面对日益增长的长文档理解需求，例如法律文书、技术规范、学术书籍等超长文本场景，仅依赖模型本身的输入 Token 容量，往往难以完整覆盖全部信息，从而限制了模型的上下文理解能力。为了解决这一问题，近年来研究者广泛采用了检索增强生成（Retrieval-Augmented Generation，RAG）[54]的范式，用以拓展模型的知识获取范围，并提升其在长文本推理与问答任务中的性能。

RAG 系统通常由三个主要模块组成：**检索（Retrieval）**、**增强（Augmentation）**与**生成（Generation）**，其核心思想如图 4.3所示。该框架通过外部知识的动态检索与有针对性的提示构建，使得语言模型能够专注于处理与问题高度相关的局部上下文信息，从而在输入限制下实现对长文本的有效理解与生成。

检索阶段：

该阶段主要针对用户输入的问题，从海量外部语料中筛选与问题高度相关的文本片段。常用的方法包括基于稠密向量的语义检索，例如使用 BERT、DPR 等编码器构建文档索引，并根据与问题的语义相似度选出 Top-K 候选段落（通常为 50 或 100

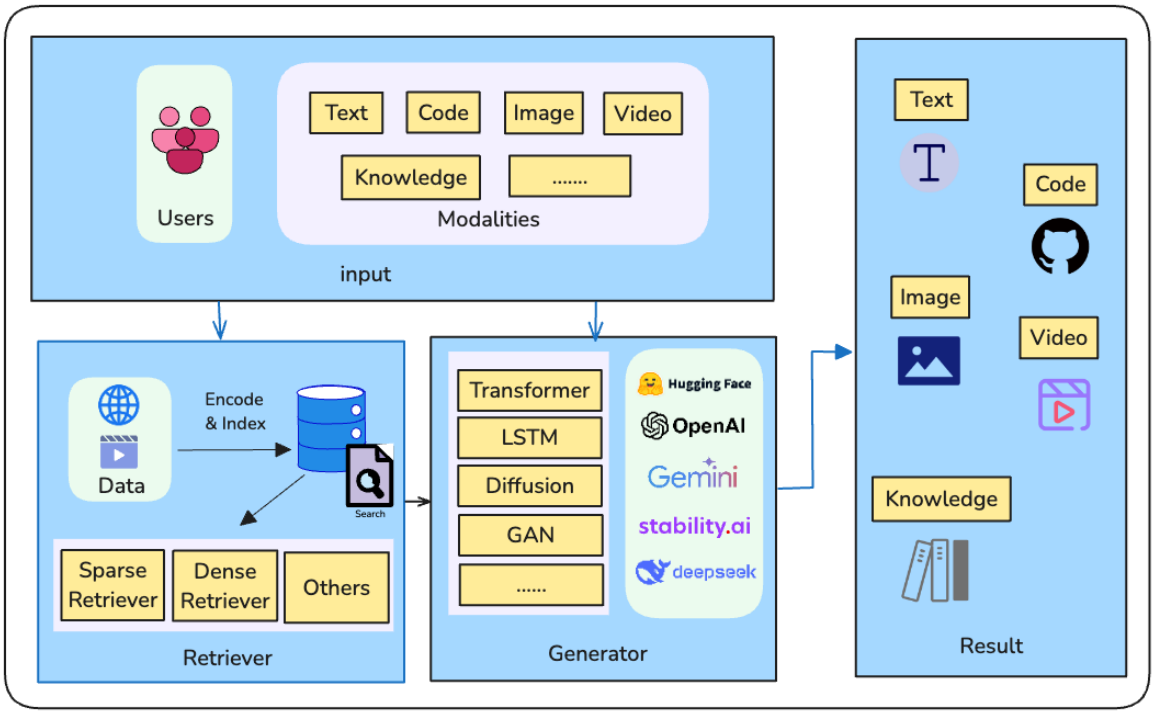


图 2.7　RAG 系统基本架构

条）。这一过程大幅压缩了原始文本的规模，使得下游模型仅需关注最具信息价值的内容。

增强阶段：

为了进一步激发大语言模型的推理与生成能力，增强阶段将检索得到的相关文段与原始问题共同组织成结构化的提示，通过合理的 Prompt 设计（例如将相关文本按特定格式拼接或分类引导）为 LLMs 提供更加明确的上下文信息。这一阶段对于最终生成结果的准确性与相关性起到了关键作用。

生成阶段：

在获取了经过精炼的信息提示后，LLMs 基于输入的 Prompt 生成最终回答或内容。得益于前两个阶段的上下文压缩与语义聚焦，生成结果不仅具备良好的语言流畅性，还在内容层面保持了较高的准确性与问题相关性，广泛应用于问答系统、对话系统、长文档摘要等任务。

综上，RAG 系统作为连接外部知识与生成模型的桥梁，在长文本理解任务中发挥着重要作用。相关研究不断探索更高效的检索方法、更合理的提示构造策略以及与生成模型的耦合机制，以进一步提升该框架在实际应用中的表现。随着文档智能研究的不断深入，从传统的单页文档分析向更复杂的多页文档理解任务扩展已成为趋势。

然而，当前在多页场景下的标注数据资源依然较为稀缺，仅有少数几个数据集在任务类型、文档长度和注释粒度方面具备代表性。这些数据集主要包括 LongDocURL、M6Doc 以及 DocVQA 中的多页理解子任务，它们分别覆盖了从长文档、中等长度文档到短页文档的不同场景，支撑了文档问答（Document Question Answer，DQA）、结构化信息提取（Structured Entity Recognition，SER）、跨文档推理等下游应用。

2.4 多页文档理解数据集研究

LongDocURL 数据集

LongDocURL[55]是当前最具代表性的超长文档理解数据集之一。该数据集包含约 50–150 页的连续文档，并围绕三类关键任务进行构建：文档理解（Document Comprehension）、跨文档元素定位（Cross-document Element Linking）和数值推理（Numerical Reasoning）。其数据由 21 位标注人员手工构建，且由 5 位硕博研究者全流程质量监督，确保了高质量的任务标注。实验表明，人类在此任务中可以轻松获得高分，但当前仅有 GPT-4o 等少数大模型能在多个任务中表现出接近人类的水平，因此该数据集被广泛用于评估多模态大模型在超长文档理解方面的能力。

M6Doc 数据集

M6Doc[56]是华南理工大学深度学习与视觉计算实验室发布的大规模中等长度文档数据集，包含 9,080 张现代文档图像，涵盖科学文章、教科书、书籍等七大类别。该数据集支持拍照文档、扫描文档和 PDF 三种格式，累计提供超过 23 万个标注实例。其丰富的文档类型和层级结构使其适合用于结构化信息提取（SER）、版面分析、以及跨文档迁移学习任务的评估。特别是在多粒度的字段抽取（如粗粒度的段落级和细粒度的字段级）任务中，M6Doc 为模型泛化能力的验证提供了良好的基准。

DocVQA 数据集

DocVQA 由微软团队构建，旨在推动视觉问答任务向文档场景拓展，强调“目的驱动”的内容理解。数据集中包含约 12,000 张文档图像和超过 50,000 个人工标注的问题答案对，文档图像主要来自 UCSF 行业文档库。DocVQA 任务不仅要求模型具备视觉文本识别能力，还需具备语义整合与跨字段推理能力。虽然原始 DocVQA 更多针对单页文档，但部分子任务已开始涉及多页上下文的建模需求，因此被广泛用作评估模型对文档语义理解能力的标准数据集之一。

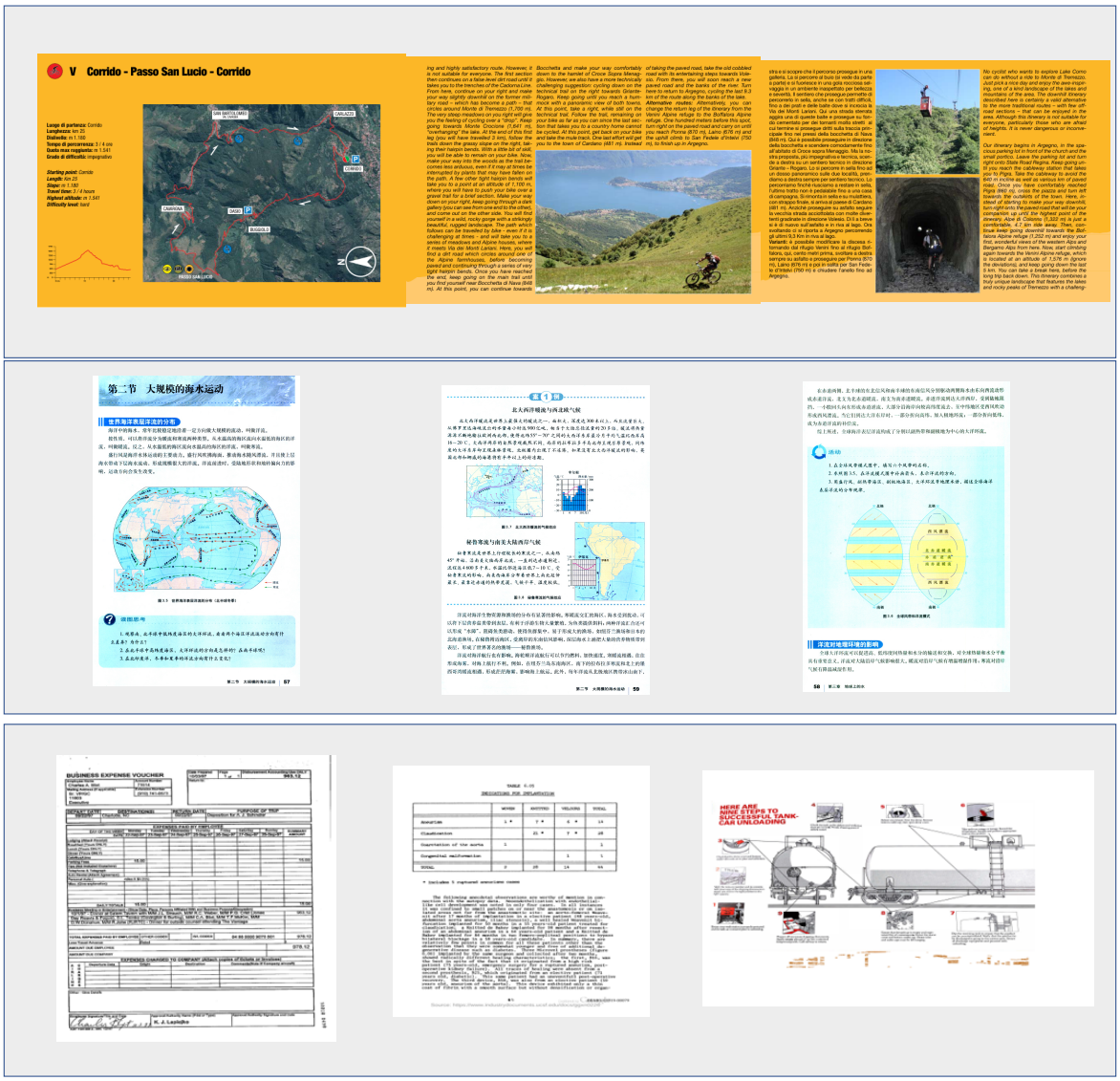


图 2.8 三种数据集的可视化，第一行是 LongDocURL，第二行是 M6Doc，第三行是 DocVQA

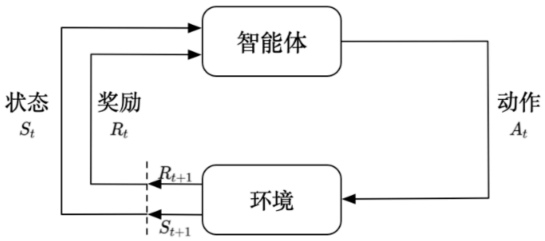


图 2.9　强化学习的交互过程

总体而言，尽管现有多页文档理解数据集已覆盖了若干代表性场景，但仍存在文档跨度有限、领域覆盖不足、以及缺乏高质量阅读顺序或人类认知标注等问题。这些挑战为后续的多页结构化理解研究与大模型评估提供了动因和研究空间。为此，构建面向真实场景、具备跨页阅读顺序标注和多任务对齐特性的长文档数据集，正逐渐成为研究热点。

2.5　强化学习

在强化学习（Reinforcement Learning，RL）领域，强化学习是通过与环境不断交互进行试错，从而学到如何优化决策的方法，如图2.9所示，整个部分由两部分组成，分别是外部的环境交互和智能体的决策[57]。整体的流程就是智能体根据当前的状态，选择采取的动作，当环境感知到选择的动作后，会流转到下一个状态，形成状态-动作-状态的流转，并且每次的状态流转都会给予奖励，智能体的目标就是将奖励累积最大化，而不是即时的奖励。

2.5.1　马尔可夫过程

马尔可夫过程是强化学习的重要组成部分，他描述了一个随机的过程，即环境的状态转移只与当前状态有关，和过去的状态无关，离散的马尔可夫过程又可以叫做马尔可夫链[58]。即在马尔可夫的时间过程中，未来状态 s_{t+1} 仅仅和当前状态 s_t 有关，不依赖于过去的状态 h_t。用公式表示即为:

$$p(s_{t+1}|h_t) = p(s_{t+1}|s_t) \tag{2.1}$$

这里的 p 指的是状态转移的概率，这个公式的含义是，系统的状态不依赖于过去的历史状态，仅依赖于当前的状态，这种不依赖过去历史状态的特性是马尔可夫过程的关键。

2.5.2 马尔可夫奖励过程

根据马尔可夫过程的特性，强化学习在每次状态转移的过程中增加了奖励的过程，所以产生了马尔可夫奖励过程，马尔可夫奖励过程是通过奖励函数，对每一个状态设定奖励，一个马尔可夫奖励过程可以表示为

$$MR = (S, P, R, \gamma) \tag{2.2}$$

每个字段的含义为:

S 表示所有可能的状态取值，P 表示状态转移的概率矩阵，这个转移概率符合马尔可夫性质，即只与当前状态有关，R 表示每个状态的奖励，γ 表示折扣因子，即智能体需要最大化累积奖励，但是排在最后一些状态的奖励是延迟的奖励，智能体会弱化延迟的奖励数值，让延迟的奖励乘以 γ 的比例和当前的奖励进行对比。因为长期的奖励具有更加的不确定性，所以需要降低延后奖励的比例，让模型更加稳定。在马尔科夫奖励过程中，我们会区分两种概念，首先是累积奖励 G 和当前奖励 R，智能体的目的就是最大化回报数值

$$G_t = R_t + \gamma R_{t+1} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \tag{2.3}$$

对于每个状态我们都用价值函数 $V(s)$ 来表示每个状态的奖励值，即当前状态奖励的期望值，公式如下

$$V(s) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | s_t = s] \tag{2.4}$$

这个公式表示，我们可以用 V 来表示当前状态 s 的好与坏，价值函数越大，代表状态越优秀。

贝尔曼方程可以用来计算马尔可夫奖励过程，贝尔曼方程如下

$$V(s) = R(s) + \gamma \sum_{s'} P(s'|s) V(s') \tag{2.5}$$

其中，$R(s)$ 是状态 s 的即时奖励，$P(s'|s)$ 是从状态 s 转移到状态 s' 的概率，$V(s')$ 是状态 s' 的价值函数。贝尔曼方程表明，某一状态的价值等于该状态的即时奖励加上其转移到其他状态后的加权期望回报。

2.5.3 策略优化原理

强化学习的目标是通过智能体与环境的交互，让智能体通过环境的反馈从而学到一个策略，这个策略可以让智能体动态的根据不同的状态，决定每个状态下所采用的动作，从而实现最大化的长期奖励，即最大化每个状态的期望回报。策略一般用 π 表示，这个策略的含义代表在给定状态 s 的情况下，智能体采取动作 a 的概率，策略包括确定性策略和不确定性策略。确定性策略为 $\pi(s)=a$，即在给定状态 s 的条件下，智能体一定会采取动作 a，非确定性策略为 $\pi(s)=\mathbb{P}(a_t=a|s_t=s)$ 即在给定状态 s_t 时，智能体按照一定概率采取某动作。

在 MDP 中，强化学习智能体会按照策略遵循策略 π，来获得最大的回报期望

$$V^{\pi}(s)=\mathbb{E}_{\pi,P}[\sum_{k=0}^{\infty}\gamma^{k}R_{t+k}|s_t=s] \tag{2.6}$$

我们通过递归计算即可获得贝尔曼方程

$$V^{\pi}(s)=\mathbb{E}_{a\sim\pi,s'\sim P}[R(s,a)+\gamma V^{\pi}(s^{'})] \tag{2.7}$$

动作价值函数 $Q^{\pi}(s,a)$ 表示在给定状态 s 和动作 a 下，使用策略 π 可以获得的回报的期望值。

$$Q_{\pi}(s,a)=\mathbb{E}\sum_{k=0}^{\infty}\gamma^{k}R_{t+k}|s_t=s,a_t=a,\pi| \tag{2.8}$$

状态的价值和动作的价值之间存在关系

$$V^{\pi}(s)=\sum_{a}\pi(a|s)Q^{\pi}(s,a) \tag{2.9}$$

表示，给定一个状态 s，他的价值等于所有可能的动作 a 的加权期望，其中每一个期望的权值就是采取动作的概率。

另一方面，动作价值可以由状态价值和奖励函数来表示

$$Q_{\pi}(s,a)=R(s,a)+\gamma\mathbb{E}_{s'\sim P(.|s,a)[V^{\pi}(s')}} \tag{2.10}$$

强化学习的拟合最终目标就是寻找一个策略 π^{*}，让智能体可以从任意时刻 t 开始，最

大化累积奖励

$$\pi^* = \textbf{argmax}\mathbb{E}[G_t|s_t = s, \pi] = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \tag{2.11}$$

对于状态价值函数而言，这个策略可以在任意时刻，任意状态 $\forall s \in \mathbb{S}$ 让状态价值函数达到最大化

$$V^{\pi *}(s) = \max(V^{\pi}(s)) \forall s \in \mathbb{S} \tag{2.12}$$

这个的含义是，在任意状态 s 开始，智能体应用最佳策略，获得的期望回报都是最大的。

在最佳策略下，智能体会选择让动作价值达到最大的动作

$$Q_{\pi^*}(s, a) = \max(Q^{\pi}(s, a)), \forall s \in S, a \in A \tag{2.13}$$

根据公式，任意状态下，最佳策略都会选择一个动作 a，使得动作价值最大，选择动作 a 之后，状态会进行转移，从而生成一条状态转移链条。所以最佳策略满足公式

$$Q_{\pi^*}(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(.|s,a)[V^{\pi *}(s')} \tag{2.14}$$

所以综上所述，我们可以得到贝尔曼最有性方程

$$Q_{\pi^*}(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[R(s, a) + \gamma \max_{a'} Q_*(s', a') \right] \tag{2.15}$$

2.5.4 强化学习策略优化方法

强化学习的核心目标是在马尔可夫决策过程（Markov Decision Process，MDP）中求解最优策略 π^*，使得从任意初始状态 $s \in \mathcal{S}$ 出发，智能体执行该策略的期望累积回报最大化。为实现这一目标，学界目前主要发展了两类方法：

- 基于价值函数的策略优化：通过迭代更新状态-动作值函数逼近最优策略；
- 基于策略梯度的策略优化：直接对参数化策略进行梯度上升优化；

我们的方法主要就是使用到了基于策略梯度的优化方法，在我们的基于下游任务自主探究阅读顺序的过程中，我们使用了策略梯度算法对阅读顺序进行了更新，之后我们又根据人类的阅读标注数据对大模型进行了广义拒绝式偏好优化（Generalized Rejection-based Preference Optimization，GRPO）微调，这些都是策略式的更新

方式。[59-61]

对于高维或连续的动作空间，基于价值函数的方法往往难以获得最优策略。另一种经典的强化学习方法是基于策略函数的方法。基于策略函数的方法的核心思想是首先构建一个策略函数，将策略参数化。设定 θ 为策略函数 π 的参数，策略 π_θ 是一个处处可微的策略，可以通过一个神经网络来建模。对于确定性策略，给定状态 s，策略输出一个确定性的动作；而对于随机性策略，则输出一个动作的概率分布。

强化学习的目标是最大化预期的折扣奖励，我们可以将策略函数的目标函数定义为:

$$J(\theta) = \mathbb{E}_{s_0 \sim \mu} \left[\sum_{t=0}^{\infty} \gamma^t R_t \right] \tag{2.16}$$

其中，$J(\theta)$ 表示给定策略参数 θ 时，智能体从初始状态 s_0 开始，经过多个时间步所获得的累积折扣奖励的期望，μ 表示初始状态分布。通过调整策略参数 θ，我们可以增大目标函数 $J(\theta)$，使得策略 π_θ 不断逼近最优策略 π^*。

因此，基于梯度上升方法，我们可以最大化期望奖励。根据公式，对于目标函数可以写为:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [G_t] \tag{2.17}$$

其中，τ 表示在策略 π_θ 下采样得到的轨迹。对目标函数求梯度的过程可以表示为:

$$\nabla_\theta J(\theta) = \sum_{\tau} G_t \nabla_\theta P_\theta(\tau) \tag{2.18}$$

其中，$P_\theta(\tau)$ 表示在策略 π_θ 下获取轨迹 τ 的概率，只有 $P_\theta(\tau)$ 与策略函数参数 θ 相关。$P_\theta(\tau)$ 可以表示为:

$$P_\theta(\tau) = \rho_0(s_0) \prod_{t=0}^{T} \mathcal{P}(s_{t+1}|s_t, a_t) \pi_\theta(a_t|s_t) \tag{2.19}$$

式中，$\rho_0(s_0)$ 表示初始状态 s_0 服从的概率分布。可以通过对 $P_\theta(\tau)$ 的对数求导获得 $P_\theta(\tau)$ 的梯度：

$$\nabla_\theta P_\theta(\tau) = P_\theta(\tau)\nabla_\theta \log \pi_\theta(a_t|s_t) \tag{2.20}$$

$$= P_\theta(\tau) \sum_{t=0}^{T} \nabla_\theta \log_{\pi_\theta}(a_t|s_t) \tag{2.21}$$

代入式 2.21 可得：

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log_{\pi_\theta}(a_t|s_t) G_t\right] \tag{2.22}$$

根据公式 2.22，我们可以直观地理解策略优化过程。在采样到的轨迹中，对于某一时刻的状态 s_t 和执行的动作 a_t，如果该动作导致最终的奖励为正，我们希望增加在该状态下选择该动作的概率。相反，如果该动作带来的奖励是负的，我们则需要减少执行该动作的概率。

具体而言，我们通过梯度上升方法来更新策略的参数。假设当前策略的参数为 θ，我们通过计算梯度 $\nabla_\theta J(\theta)$ 来更新 θ。更新规则如下：

$$\theta \leftarrow \theta + \eta \nabla_\theta J(\theta) \tag{2.23}$$

其中，η 是学习率，控制着每次更新的步长。

在策略优化过程中，智能体不断与环境交互，执行动作并收集反馈数据。在每次与环境交互时，智能体都根据当前的策略选择动作，直到任务完成。通过这种方式，智能体可以生成多条轨迹，每条轨迹中包含一系列的状态和动作对，以及这些轨迹下对应的一整条奖励信号轨迹。我们可以通过不断收集这些数据，使用蒙特卡洛的方法采样样本轨迹来优化计算出对应的策略梯度，并将其代入公式 2.22 来不断优化策略。

基于这个方法，智能体会在多次迭代过程中逐步改进策略，通过梯度上升不断调整策略参数，使得期望的累积奖励逐渐增加，最终逼近最优策略 π^*。

2.5.5 PPO 算法

近端策略优化（Proximal Policy Optimization，PPO）[62-63]是一种强化学习中的策略优化方法，属于基于策略（Policy-based）的算法。PPO 通过优化一个代理目标函数来提高策略，同时限制新旧策略之间的更新幅度，从而在保证学习效率的同时增强训练稳定性。

与信赖域策略优化（Trust Region Policy Optimization，TRPO）相比，PPO 不需要计算复杂的 Fisher 信息矩阵或进行约束优化，而是采用更简单的技巧——**剪切（Clipping）**或 **KL 惩罚**，来限制策略更新的“接近程度（Proximity）”。

2.5.5.1 数学公式与目标函数

PPO 的核心在于优化以下的 **剪切目标函数（Clipped Surrogate Objective）**：

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \tag{2.24}$$

其中：

- $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ 表示当前策略与旧策略在动作 a_t 下的比值。
- $\hat{A}_t$ 是优势函数（Advantage Function）的估计，衡量动作 a_t 相较于平均策略的好坏。
- ϵ 是一个超参数，控制更新的范围，一般取 0.1 到 0.3。
- clip 函数将比值 r_t 限制在 $[1-\epsilon, 1+\epsilon]$ 区间。

这种形式可以防止策略在某些步骤上更新过快，避免导致策略崩溃（catastrophic collapse）。

2.5.5.2 优势函数估计

通常，PPO 使用广义优势估计（Generalized Advantage Estimation, GAE）来计算 $\hat{A}_t$，形式如下：

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l} \tag{2.25}$$

其中：

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{2.26}$$

- γ 是折扣因子；
- λ 是 GAE 中控制 bias-variance 权衡的参数；
- $V(s_t)$ 是状态值函数估计。

2.5.5.3 总体损失函数

实际训练中，PPO 的总损失函数通常包含以下三部分：

$$L^{\text{PPO}} = \mathbb{E}_t \left[L_t^{\text{CLIP}}(\theta) - c_1 \cdot (V_\theta(s_t) - R_t)^2 + c_2 \cdot \text{Entropy}[\pi_\theta](s_t) \right] \tag{2.27}$$

其中：

- 第一项是策略目标；
- 第二项是值函数的均方误差（用于训练 critic）；
- 第三项是策略熵，用于鼓励探索；
- c_1，c_2 是超参数。

所以 PPO 算法被广泛应用于近年来的大模型训练调整中，但是 PPO 算法存在着训练奖励函数过程过于复杂的问题，所以针对于以上的问题。DeepSeek 团队提出了 GRPO 算法来解决下面的问题。

2.5.6 GRPO 的目标函数推导

GRPO 试图最大化以下目标函数，该目标可被视作偏好排序下的一种近似对比损失：

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \left(\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_w|x) + \pi_\theta(y_l|x)} \right) \right] \tag{2.28}$$

这是一个 **对比性 softmax** 损失，其思想类似于将偏好判断视为一个二分类任务。具体解释如下：

- 如果策略 π_θ 对 y_w 的概率远大于 y_l，则损失较小，说明模型与偏好一致。

- 如果策略对 y_l 的概率较高，则损失增大，模型将受到梯度惩罚以向更偏好的方向调整。

这一形式与人类自然判断偏好一致，也方便梯度计算和稳定训练。从一个组内挑选更优秀的内容参与迭代控制。

第三章　基于智能体自主探索的单页文档阅读顺序生成方法

3.1　研究动机

准确抽取文档的阅读顺序是文档智能领域的一项基础且关键的任务。现有主流方法通常依赖于与人类阅读顺序严格对齐的标注数据进行智能体训练。这类标注数据在指导模型学习常规布局的阅读路径方面取得了一定成效，并且高质量的阅读顺序信息已被证明能显著提升语言模型对文档语义结构的理解能力。然而，随着视觉富文本文档的广泛应用，文档布局呈现出前所未有的多样性和复杂性（如多栏、嵌套表格、不规则文本框、图文混排等），对阅读顺序的鲁棒性提出了严峻挑战。在此背景下，依赖人类标注的阅读顺序面临显著瓶颈:

- **标注成本高昂且可扩展性差:** 获取高质量、覆盖广泛布局类型的人类阅读顺序标注需要耗费巨大的人力物力，尤其对于复杂、非标准化的文档，标注过程本身也充满主观性和不确定性。
- **泛化能力受限:** 人类标注的数据集难以穷尽现实世界中所有可能的布局变体，导致基于这些数据训练的模型在面对训练集分布之外的、特别是高度复杂或新颖的布局时，其阅读顺序预测的鲁棒性急剧下降。

作为替代方案，基于规则（如严格遵循坐标信息的“从左到右、从上到下”）的解析方法虽然无需标注，但其固有的僵化特性使其更难以有效适配千变万化的文档布局。这类方法往往无法处理常见的布局复杂性问题（如多栏文本流的切换、文本绕排、跨页内容等），其解析结果常与实际的逻辑阅读顺序存在偏差。

因此，当前主流的阅读顺序抽取模型（无论是基于标注学习还是基于规则）普遍存在一个根本性缺陷: 它们本质上是对特定模式或有限布局规则的学习或模仿，其行为被严格约束在训练数据或预设规则所定义的边界内。这些模型缺乏真正的布局理解能力和推理灵活性，无法像人类读者那样根据上下文语义和视觉线索，动态地、创造性地“打破”固有模式，适应全新的、未曾见过的布局结构。

为了解决上述限制，深入探索语言模型在处理下游文档理解任务（如信息抽取、问答、摘要）时真正需要且最有效的阅读顺序，并彻底摆脱对大规模、高质量人工标

注数据的强依赖性，本章研究致力于解决以下两个核心科学问题:

- 如何设计一种阅读顺序抽取机制，使其能够**超越对固定布局模式的依赖**，具备强大的鲁棒性和泛化能力，以应对高度复杂且多变的 VRD 布局?
- 如何在不依赖昂贵人工阅读顺序标注的前提下，有效驱动模型学习到符合文档内在逻辑的、**可服务于下游任务**的阅读顺序?

3.2 任务定义

在文档理解任务中，**阅读顺序的抽取（Reading Order Extraction）**可形式化地视为一个**离散的排列组合问题**。具体而言，给定一张文档图像 $\mathcal{D}$，可通过 OCR 引擎提取出其初始的文字内容序列 $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$ 及对应的位置信息 $\mathcal{B} = \{b_1, b_2, \ldots\ b_N\}$。由于 OCR 的输出顺序往往与人类的实际阅读顺序不一致，因此需要对该文字-位置对 $(\mathcal{T}, \mathcal{B})$ 进行重排序，以更符合自然的阅读习惯和文档结构逻辑。

在完成排序后，重排后的文本及其布局信息可作为输入喂入预训练语言模型，从而执行一系列下游任务，如**结构化信息抽取**和**文档视觉问答（Document Visual Question Answering，VQA）**等。

对于 SER 任务，其目标是根据文字内容及其对应的几何布局信息，识别每个文字片段所属的实体类别。形式上，该任务可表示为一个映射函数:

$$F_{IE}(\mathcal{D} : \langle \mathcal{B}, \mathcal{T} \rangle) \rightarrow \epsilon \tag{3.1}$$

其中，ϵ 表示每个文本块的预测标签，标签遵循标准的 BIO 编码方案（即 {Begin Inside Outside}），并从一个预定义的实体类别集合中选取。

对于 VQA 任务，其形式定义为: 在给定文档图像 $\mathcal{D}$ 和自然语言问题 $\mathcal{Q}$ 的条件下，语言模型需要生成一个符合语义的回答，即:

$$Answer = LM(\mathcal{D}, \mathcal{Q}) \tag{3.2}$$

该任务评估模型对文档图像中结构化与非结构化信息的理解能力，要求模型能够综合利用视觉布局、文本内容与问题语义之间的关系进行推理与生成。

综上，阅读顺序的合理抽取不仅有助于恢复文档原有的语义结构，也显著提升了文档理解模型在各类下游任务中的性能表现。

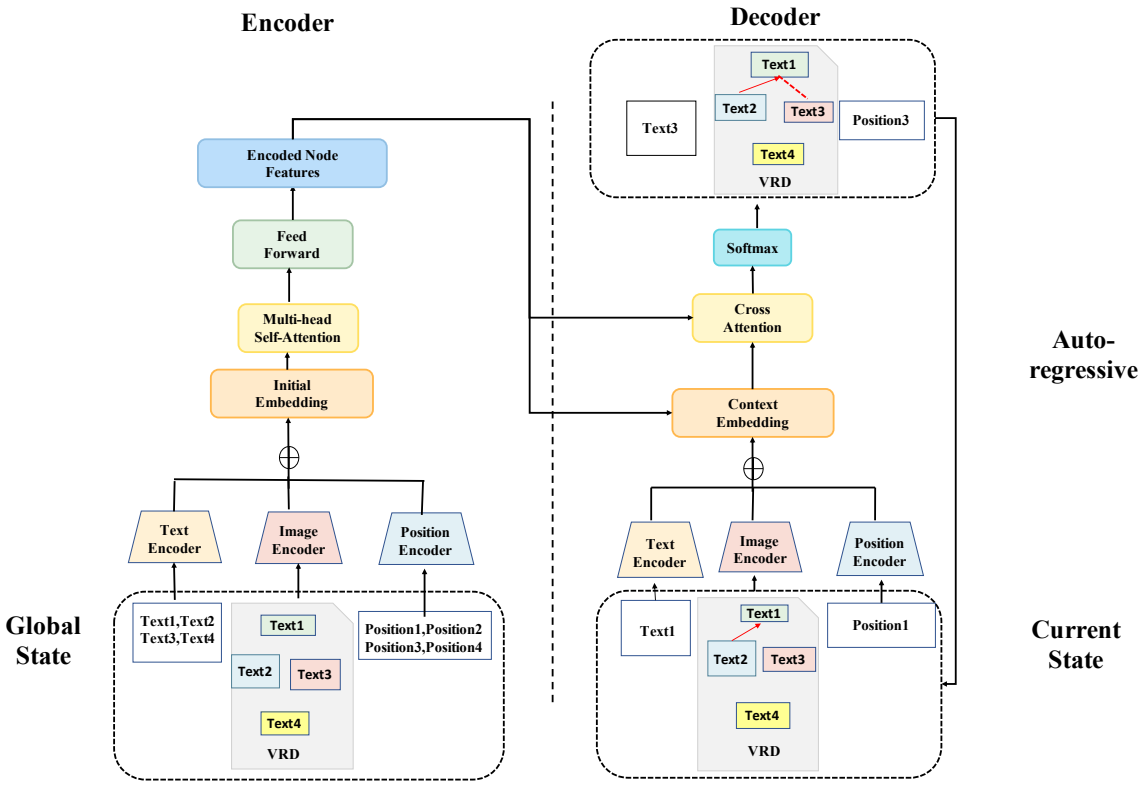


图 3.1　模型架构原理

3.3　提出模型

为了使阅读智能体能够自主完成整个阅读过程，并探索出一条不同于传统人类阅读顺序的路径，即打破人类在阅读中通常遵循的线性或默认规则所限定的先后顺序，本文受到近年来强化学习相关研究成果的启发，将阅读任务抽象为一个组合优化问题，提出了一种新颖的阅读顺序抽取方法，称为基于人类启发的阅读智能体（Human-Inspired Reader，简称 HiDReader）。

该方法通过模拟人类在自然阅读过程中的认知机制[64]，将阅读行为建模为**局部阅读**与**全局阅读**的协同过程。在实际的阅读行为中，人类通常不仅关注当前阅读的词语或句子，还能够从整体排版、段落结构、标题层级等全局信息中获得有助于理解的线索。同时，人类能够记忆和整合已阅读的内容，并据此对未阅读的信息进行推理与预测。因此，我们将这种认知过程引入到智能体的阅读策略设计中，将文本建模为由局部信息与全局信息共同组成的表示空间。

在模型执行阅读决策的每一步，智能体都会基于当前已阅读的内容（作为局部上下文）以及整个文档的全局布局与语义表示，通过注意力机制来评估下一个最优的阅读位置，即待阅读的 Token。该机制不仅提升了阅读过程的动态适应性，也为阅

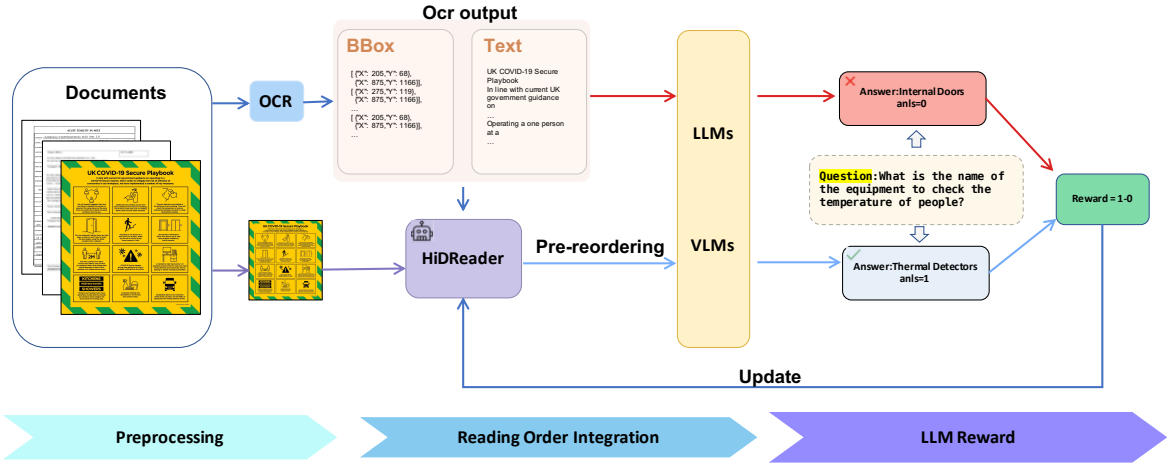


图 3.2 整体的处理框架

读路径的个性化建构提供了理论支撑。

HiDReader 在训练过程中采用了强化学习框架，将整个阅读路径视为一次状态-动作序列，通过设定下游任务的整体性能指标（如信息提取、文本分类等）作为最终奖励信号，优化智能体对阅读顺序的学习策略。该方法充分发挥了 Transformer 架构在建模长距离依赖关系方面的优势，结合局部与全局信息之间的相互作用，有效提升了阅读策略对下游任务的适应性与泛化能力。从而可以很好的获取到有利于下游任务的阅读顺序。

3.3.1 整体框架

HiDReader 包含三个部分，分别是编码器部分、解码器部分和奖励调节参数部分。编码器部分负责提取和融合多种模态的特征，包括图片、文字、坐标模态。解码器部分我们采用 Transformer 进行解码，通过 Transformer 强大的全局注意力捕捉能力进行阅读顺序的解码过程。我们首先对文字模态，图片模态和坐标模态分别进行编码。文字模态我们采用预训练后的 BERT 的编码器进行编码，图片模态我们使用 ViT[65] 进行编码，坐标模态我们使用图神经网络进行编码。

$$E_T = BERT(Text) \tag{3.3}$$

$$E_I = ViT(Image) \tag{3.4}$$

$$E_B = GraphEmbedding(Bbox) \tag{3.5}$$

最后我们对三种模态进行融合，获得最终的模态表示。

$$E_{T+B+I} = E_T + E_B + E_I \tag{3.6}$$

我们训练的目标是最大化下游任务的表现，具体包括 SER 任务的指标 F1 值和 VQA 任务的指标 ANLS。F1 值的计算公式见3.26。这里的 $Precision$ 表示预测为正例的真正的正例的占比，$Recall$ 表示为真正的正例中，被预测的正例占比，所以 F1 值可以很好的衡量语言模型的表现能力，不会因为正负例样本的不均衡导致衡量的性能有偏差。F1 的取值在 0-1 之间，越接近 1，说明模型的能力越好。ANLS 分数是用来衡量 VQA 准确度的指标，计算公式见3.28 其中 $Levenshtein$ 距离代表从字符串 a_i 到 p_i 的最小变化次数。通过 ANLS 的指标，可以很好的衡量 VQA 的模型预测回答与真实回答内容之间的差距。

3.3.2 基线对抗的奖励设计

我们的目标是重新排列输入到语言模型的 OCR 结果，给定一组 OCR 解析的默认结果 $T_d = \{t_1, ...t_n\}$，将这组 OCR 解析得到的默认阅读顺序输入到下游语言模型中，例如我们使用了预训练的小模型，例如 BERT 纯文本模态的模型，LayoutLM 系列多模态预训练模型，以及现在非常火爆的大型语言模型，同样我们也分为了纯文本大语言模型和多模态大语言模型，纯文本大语言模型包括 Qwen1.5-32B[66]，LLaMa3-8B 模型，多模态大语言模型包括 Qwen2VL-7B[67]，GLM4V-9B[68]，通过多种不同模态的模型对我们的阅读顺序挖掘算法进行验证。经过我们 HiDReader 重新排列的阅读顺序为 $T_h = \{t_{r_1}, t_{r_2}, ... t_{r_n}\}$ 我们的奖励函数设计如下

$$R_m = \begin{cases} F_1(T_h) - F_1 T_d & \text{if } task = SER \\ ANLS(T_h) - ANLS(T_d) & \text{if } task = VQA \end{cases} \tag{3.7}$$

这个奖励就是我们的主要奖励，具体的作用就是通过最大化重排序之后的文本和默认顺序文本在下游任务中的表现，从而指导智能体进行下游任务的进行。

3.3.3 保持语义稳定的奖励

为了保证我们的语义信息可以在重排序之后尽量的保持，我们需要让重排序之后的语义信息和默认顺序的语义信息进行一定程度上的对齐，这样可以增强强化学

习过程中的稳定性，所以我们使用了相似性函数来衡量两个文字之间的相似度。表示为公式

$$R_s = sim(T_r, T_o) \tag{3.8}$$

- R_s 表示语义相似度奖励
- T_r 表示重排列顺序后的文本
- T_o 表示原始文本

最终，我们采用混合奖励的方式作为我们的最终奖励，分别对语义奖励和下游任务的奖励进行加权求和

$$R_f = \lambda \times R_m + (1 - \lambda) \times R_s \tag{3.9}$$

这里的奖励就是将两种奖励进行求和，从而指导智能体最终可以更加稳定的获得让下游任务表现更优的阅读顺序，在这里我们的 λ 取值为 0.8，含义就是比例参数。通过这样的奖励设计，我们将默认的阅读顺序作为基准阅读顺序，我们优化的目标是最大化奖励，通过这样的奖励设计，可以让强化学习智能体可以感知到如果奖励为正，则代表本次阅读顺序更优，如果奖励为负，则代表本次阅读顺序更差。所以这样的奖励设计，可以让智能体进行动态调整策略，让智能体朝着更优的方向进行优化。并且我们可以通过语义信息来稳固智能体进行强化学习的过程，让智能体在学习初期可以充分理解语序和目标文字之间的差异性。

3.3.4 局部与整体注意力计算

人类阅读的过程往往包括两个阶段，首先我们看到一个文档时，会浏览到整体的排版布局，所以我们会拥有文档整体布局的先验知识，其次，我们的阅读往往会通过文档中的视觉引导进行阅读，并且通过局部信息在整体中的位置来决定下一次阅读的焦点位置。所以人类的阅读会使用到文字，坐标布局，文档图片三种模态。并且根据这三种模态的信息和整体的布局进行下一次阅读位置的决策。所以我们受到人类阅读行为的启发进行阅读顺序的设计，利用融合了三种模态的信息进行局部编码的表示，将局部编码表示为 E_{T+B+I}，其中 T 表示当前阅读的文字信息，B 表示当前阅读的坐标信息，I 表示在图片中显式标注的文字信息。并且将整体的多模态信息

通过 Transformer 进行特征的提取，表示为

$$E_{total} = Attention(Q, K, V) \tag{3.10}$$

这里的 Q, K, V 都是所有的 Token 的特征表示，通过自注意力计算得到更加精准的编码表示。接下来，我们通过局部和整体的注意力计算得到通过当前 Token 来决定得到下一个阅读位置的计算。公式为

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{3.11}$$

这里的 Q 为局部的编码表示，K，V 为全局的编码表示。

$$ContextEmbedding = Attention \odot Attention_w \tag{3.12}$$

这里的 $Attention$ 是3.10，$Attention_w$ 指的是注意力的权重信息通过注意力和注意力权重相乘，获得上下文向量。每一个纬度都对应一个待选择的文字，为了避免我们的智能体重复选择某一个文字，我们引入了 $Mask$[69] 对每一个文字进行掩码建模

$$Mask = \begin{cases} 0 & \textbf{if } pick = True \\ 1 & \textbf{if } pick = False \end{cases} \tag{3.13}$$

这里的 $pick$ 指的是当前的文字是否被选择，如果已经选过了则 Mask 为 0，未选择则 Mask 为 1。最终的每个位置的选择，通过公式

$$P_t = Softmax(ContextEmbedding * Mask) \tag{3.14}$$

$$\hat{y}_t \sim Bernoulli(P_t) \tag{3.15}$$

这里的 $Bernoulli$ 指的是伯努利采样，最终的概率通过伯努利采样获得下一次需要阅读的文字索引。

3.3.5 策略梯度更新

由于阅读顺序抽取任务可以被看作一个马尔可夫过程，给定状态表示 $state$: $\{T_r, T_t\}$，这里的 T_r 表示已阅读的文本，T_t 表示所有的文本。动作空间的表示为 A :

$\{0, 1, ...N\}$, N 表示文本的总数, 所以这里的含义为每次采样一个待阅读的词语, 我们的状态都会通过多模态融合表示为编码，初始状态我们定义为一个取值范围在 (0,1) 之间的均匀分布，维度为 D，这里的维度我们使用的是 300。我们对多模态融合的向量维度取为 D，初始化状态编码表示为 D 上的均匀分布。阅读顺序抽取问题可以看作是一个只有完整阅读完所有文字才可以获取最终阅读质量奖励的稀疏奖励过程, 为了稳定稀疏奖励所带来的策略不稳定的问题，我们对上述奖励函数进行了标准化和正则化。

$$R_{\text{mean}} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} R_i & \text{for the initial batch} \\ \beta R_{\text{mean}} + (1-\beta)\mathbb{E}[R] & \text{for subsequent batches} \end{cases} \tag{3.16}$$

接下来我们对奖励进行放缩，让模型可以感知到这次的采样对模型的影响是积极的还是消极的

$$A = R - R_{\text{mean}} \tag{3.17}$$

所以我们采用策略梯度的方法对神经网络模型的参数进行更新。为了可以更新具体的参数信息，我们对每个概率进行 Log 计算。

$$\log P = \sum_{i} \log P_i \tag{3.18}$$

为了避免概率值过低，我们设定了最低值-1000，防止值过低产生 NAN,

$$\log P = \max(\log P \ -1000) \tag{3.19}$$

最后我们使用策略梯度函数的更新策略进行更新。更新的目标函数如下

$$\nabla_\theta L_{\text{actor}} = \mathbb{E}[\nabla_\theta (A \cdot \log P)] \tag{3.20}$$

最后我们根据学习率 η 对模型的参数进行更新

$$\theta \leftarrow \theta - \eta \nabla_\theta L_{\text{actor}} \tag{3.21}$$

3.4 评价指标

为了进一步量化 HiDReader 所生成阅读顺序与人类阅读顺序之间的相似性，我们引入了两种常用的秩相关统计指标: 肯德尔秩相关系数（Kendall’s Tau）[70] 与斯皮尔曼等级相关系数（Spearman’s Rho）[71]，对模型在不同模态输入条件下生成的阅读序列进行了比较分析。

3.4.1 顺序相似度评估指标

在顺序相似性评估任务中，斯皮尔曼相关系数和肯德尔系数是常用的统计方法，主要用来统计两个分布的单调关系。这两种统计方法适合于处理序列顺序的数据，用于衡量序列之间的关系，无需让数据符合正态分布，主要应用场景是

1) 评估序列一致性: 计算斯皮尔曼系数和肯德尔系数可以衡量两个时间序列的相似度，阅读顺序是一个符合时间序列分布的数据。

2) 评估非线性关系的相似度: 这两种系数对非线性关系的敏感度较高，适用于分析相关数据。

3) 处理秩次数据: 当数据类似于排名数据时，使用这两个更为合适。

4) 鲁棒性: 当数据为异常数据时，这两种系数的敏感度更低。

肯德尔系数较多的被用来衡量两个序列之间的一致性，需要在衡量全局局部排序一致性的时候使用肯德尔系数，斯皮尔曼系数用来衡量两个序列之间的单调关系，可以捕捉全局的趋势。

肯德尔系数: 肯德尔系数是一种衡量两种分布之间排序一致性的统计数值，定义如下给定如下两个排列 $X = (x_1, x_2 \ldots x_n)$ 和 $Y = (y_1, y_2 \ldots y_n)$ 肯德尔系数 τ 的计算公式为

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sng(x_i - x_j) sgn(y_i - y_j) \tag{3.22}$$

其中:

1. $sgn(\cdot)$ 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
2. n 代表序列的长度
3. 肯德尔系数的取值范围为 $[-1, 1]$，$\tau = 1$ 时表示两个序列完全一致，$\tau = -1$ 时

表示两个序列完全相反，$\tau = 0$ 时表示两个序列没有相关性

2) 斯皮尔曼系数斯皮尔曼系数是衡量两个序列之间单调性的相关系数，定义如下: 给定两个序列 X 和 Y，记为 R_X 和 R_Y。斯皮尔曼系数 ρ 的计算公式为:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3.23}$$

其中:

1. $d_i = R_X(i) - R_Y(j)$ 是第 i 个元素的秩次差
2. n 是元素的长度
3. 斯皮尔曼系数的取值范围为 $[-1, 1]$，$\rho = 1$ 时表示两个序列完全一致，$\rho = -1$ 时表示两个序列完全相反，$\rho = 0$ 时表示两个序列没有相关性。统计结果如表所示。

3.4.2 命名实体识别任务评价指标

命名实体识别任务的目标是要求 LLM 理解文档的语义和位置关系，并且给该实体一个推断一个它的类型。在 LLM 的二维空间位置理解评估，在 SER 任务中存在粗粒度和细粒度的 SER。粗粒度的 SER 只关心语义实体的健和值，细粒度的 SER 任务要区分实体。

评估方法: 根据类别的不同，DIE 任务可以被认为是分类任务，所以评估方法要用分类任务的常用指标即利用基于混淆矩阵 (Confusion Matrix) 的准确性度量来衡量 LLMs 位置判断的能力。混淆矩阵展示了实际类别与模型预测类别之间的关系。以二分类问题为例，混淆矩阵如表 4.1 所示。

表 3.1　混淆矩阵

真实结果	**预测结果**	
	正类	**负类**
正类	TP（真正类）	FN（假负类）
负类	FP（假正类）	TN（真负类）

计算精确率 (Precision，P)、召回率 (Recall，R)、F1 分数 (F1-score，F1) 的计算公式如下:

(1) 精确率 (Precision，P): 精确率衡量的是被正确识别为正类的样本占模型识别

为正类的样本的比例。

$$P = \frac{TP}{TP + FP} \tag{3.24}$$

(2) 召回率 (Recall，R): 召回率衡量的是被正确识别为正类的样本占实际正类样本的比例。

$$R = \frac{TP}{TP + FN} \tag{3.25}$$

(3) F1 分数 (F1-score，F1):F1 分数是精确率和召回率的调和平均，用于在精确率和召回率之间取得平衡。

$$F1 = \frac{2 \times P \times R}{P + R} \tag{3.26}$$

3.4.3 文档问答任务评价指标

文档问答（Document QA）任务的目标是要求 LLM 能够在提供的文档中，定位并抽取出与问题最相关的内容片段作为答案。这一任务通常涉及文本的匹配与语义理解，尤其是在文档结构复杂、信息分散的场景下，模型需要具备较强的信息整合与推理能力。

评估方法: 文档问答任务常用的评估指标为平均归一化 Levenshtein 分数（Average Normalized Levenshtein Similarity，简称 ANLS）。该指标既考虑了预测答案与真实答案之间的字符级相似度，又引入了阈值以增强其区分能力。ANLS 可以有效地衡量模型输出与参考答案之间的近似程度，允许一定的字符误差或表达差异。

具体计算方法如下:

- 首先，对每一组预测答案（Pred）和真实答案（GT），计算其 Levenshtein 距离（编辑距离），记为 $Lev(Pred, GT)$。
- 然后，归一化该距离:

$$Lev_{norm} = \frac{Lev(Pred, GT)}{\max(Pred, GT)} \tag{3.27}$$

- 若归一化后的编辑距离小于阈值 $\tau = 0.5$，则本轮 ANLS 得分为 $1 - Lev_{norm}$，否则记为 0。

最终整个数据集上的平均 ANLS 得分为:

$$ANLS = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{3.28}$$

其中，$s_i = 1 - Lev_{norm_i}$ 当 $Lev_{norm_i} < \tau$，否则 $s_i = 0$，N 为总的问题数量。

该指标综合考虑了答案的语义接近性与形式误差，广泛用于 DocVQA 等文档问答评估任务中。

3.5 实验

我们在 DocTrack，DocVQA 数据集上进行了实验用来衡量经过我们的强化学习对下游任务的模型表现有怎么样的影响，所以我们采用了文档理解任务中的 SER 和 VQA 进行了我们的实验，同时衡量多种大模型的性能影响。

3.5.1 数据集介绍

实验使用了 DocTrack,DocVQA 数据集进行主实验。DocTrack 数据集是一个有人类眼动标注数据的数据集，里面包括 FUNSD[51]，Seabill[17]，InfoGraphic 三个数据集，使用 TobiiStudio 进行人类眼动信息采集。DocTrack 数据集的目的是通过采集人类阅读文档的数据来增强语言模型的理解能力。数据集包括

表 3.2　数据集内容统计

#	**Pattern**	**Funsd** norm-z	**Seabill** local priori	**InfoGraph** cross&visual	**Total** -
Train	*doc*	149	160	100	409
	ent	7441	10024	12650	30115
	tok	22512	16055	24364	62931
Test	*doc*	50	50	30	130
	ent	2332	3430	3794	9556
	tok	8973	7022	7308	23123

- **FUNSD** 数据集主要包括纯文本的文档，文档的格式内容按照 Z 字分布，阅读顺序从左到右，从上到下。FUNSD 数据集是 SER 任务，每个语义实体包括一

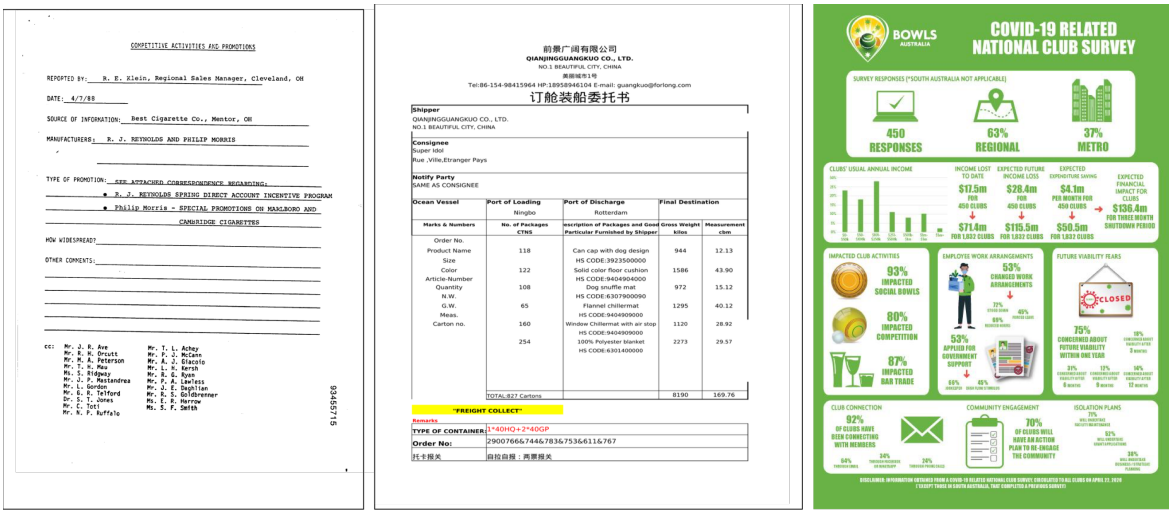


图 3.3　FUNSD（a）、Seabill（b）和 InfoGraphic（c）数据集的样例展示

个唯一标识符 id，一个语义标签 ({问题，答案，标题和其他})，一个边界框，一个实体链接列表，以及一个单词列表。

- **Seabill** 数据集是从国际海运场景中提取的表单数据，包括 3562 个训练文档和 953 个测试文档。数据由 PDF 图片和基于 PPOCRLabel 标注的文本、位置和图片组成，具有三个粗粒度的标签，{标题，问题，答案} 和 56 个细粒度的标签包括 {发货人，发货人-value，起运港，起运港-value，...}。
- **InfoGraphic** 数据集这个数据集是由 ICDAR 提出的问题中提取的图片，由人类进行视觉标注引导，包括 100 张训练集和 30 个测试集数据，数据由 Tobii Studio 标注获得，包含 Text，Bbox 信息和图片信息。每一个文档都对应一个 $\{Q, A\}$ 对。

3.5.2　骨干网络

在阅读顺序建模模块的设计上，我们借鉴了 PointerNetwork 的思想。传统的 PointerNetwork 通常采用 LSTM 作为序列编码与注意力计算的基础模块，其核心通过指针机制对输入序列中的元素进行位置选择，广泛应用于排序、路径规划等结构化输出任务中。然而，LSTM 的建模能力在处理长距离依赖关系时存在一定的局限性。

为提升模型对全局信息的捕捉能力，我们在 HiDReader 中将 LSTM 替换为结构上更具优势的 Transformer 模块。Transformer 引入了自注意力机制（Self-Attention），能够并行建模任意位置间的依赖关系，显著提升了模型的表示能力与推理效率。在

此基础上，相关工作如 PointerFormer[72] 等进一步将 Pointer Network 与 Transformer 结构融合，通过多层自注意力机制对输入特征进行细致建模，取得了优异的排序性能。

在本研究中，我们进一步扩展了 PointerNetwork 的基础结构，引入多模态信息融合机制。具体而言，我们将来自文本、图像、坐标等模态的特征编码后进行融合，以丰富节点间的语义与空间表达能力。在指针机制选择目标节点的过程中，多模态融合特征被用于计算注意力权重，从而实现更加准确的阅读顺序预测。

此外，为了提升模型在特定任务中的适应性，我们引入了基于强化学习的训练范式。与原始 PointerNetwork 相同，我们采用采样式策略网络对节点排序进行建模，但不同之处在于我们优化的目标函数不再是重构某一固定的标签顺序，而是通过最大化下游文档理解任务的性能指标（例如 F1、ANLS 等）作为强化学习中的奖励函数，指导模型学习更具任务价值的阅读顺序。这种方式实现了从任务导向的角度对排序策略进行动态优化，使得 HiDReader 能够在多模态语义建模与任务表现之间取得良好的权衡。

3.5.3 实验细节

HiDReader 使用 Adam 优化器进行优化，学习率设置为 0.9，没有预热的过程。模型的学习率设置为 1e-4，奖励设置为放缩后的奖励，并且使用 Welford 算法来进行奖励的存储，我们设置最大的文字数量为 100，训练了 1000000 个 step 让模型达到收敛，让模型可以稳定的获得超越默认阅读顺序的表现。其他实验的基本配置如下表我们使用了包括传统的语言模型包括 BERT，LayoutLMv2，LayoutLMv3，以及纯文

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	976G
CPU 处理器	Intel(R) Xeon(R) Gold 6348 CPU @2.60GHz
GPU 处理器	GeForce RTX 3090
Python 版本	2.3.0
Pytorch 版本	3.10
CUDA 版本	12.1
transformer 版本	4.5.0

本大模型 Qwen1.5-32B，LLaMa3-8B 和两种视觉大模型 GLM-4V-9B，Qwen2-VL-7B，这些模型的参数配置如下

表 3.4　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
BERT (Base)	2018-10	110M	33 亿词	512	英语	是
LayoutLMv2[5]	2020-10	200M	5M 文档	512	英语	是
LayoutLMv3[6]	2022-02	200M	11M 文档	512	英语	是
Qwen1.5-32B[66]	2024-01	32B	3 万亿 Token	32K	中、英	是
LLaMA3 8B[39]	2023-03	8B	15 万亿个 Token	8K	英语	是
GLM-4V 9B[68]	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

3.5.4　多模态阅读顺序对文档理解性能的影响分析

表 3.5-3.7 展示了在 FUNSD、Seabill、InfoGraphic 以及 DocVQA 四个基准数据集上的实验结果，具体包括精确率（Precision，P）、召回率（Recall，R）以及 F1 分数。表格中使用 **粗体**标示性能最优的结果，下划线 标示性能次优的结果，以突出不同方法之间的性能差异。

在实验设计中，我们针对三种不同的模态信息进行了消融实验，分别为文本模态（T）、坐标模态（B）以及图像模态（I）。此外，为了验证阅读顺序对文档理解任务的影响，我们引入了多种预排序策略（PREORDER）进行对比，包括: 人类真实阅读顺序（EYE）、默认的 OCR 解析顺序（DEFAULT-OCR）、严格遵循从左至右、从上至下的规则式排序（Z-ORDER）、基于局部优先思想的 XYLAYOUT 策略，以及我们提出的 HiDReader 模型基于不同模态学习得到的阅读顺序。

在实验中，我们分别采用纯文本语言模型、多模态语言模型、纯文本大模型以及多模态大模型对上述数据集进行评估，结果表明我们提出的方法具备良好的通用性。仅通过调整阅读顺序，即可显著提升各类语言模型在文档理解任务中的性能，且在所有任务中均超越默认阅读顺序（DEFAULT-OCR）所构建的基线模型。

值得注意的是，在部分任务中，Z-ORDER 阅读顺序在某些指标上表现略优于我们的方法。其原因可能在于当前主流语言模型在预训练过程中广泛使用了基于 Z-ORDER 的规则式阅读顺序作为训练信号，从而使得模型在面对此类结构时具备更强

的适应能力。

进一步分析纯文本大模型的表现可以发现，在单页文档场景中，Z 字形的阅读顺序往往更有利于模型理解。这主要归因于纯文本模型缺乏对空间结构的感知能力，难以有效捕捉文档的二维布局信息，而 Z-ORDER 排序在模型的预训练阶段频繁出现，从而使其对该排序结构具备更强的记忆偏好。因此，在此场景下，尽管我们基于强化学习的方法能够在一定程度上优化阅读顺序，但其搜索解空间仍略逊于固定的 Z-ORDER。

在多模态大模型的实验中，我们观察到对于结构相对简单的单页文档（如表单类文档），Z-ORDER 排序因其符合表格类信息从左至右、从上至下的自然排列方式，表现出良好的性能。然而，在结构更为复杂的文档类型中，如信息图（Infographic），我们的方法展现出更强的泛化能力和适应能力。这是由于信息图中的图像模态信息更为丰富，且文本与图像信息之间往往存在复杂的语义关系。通过多模态融合及强化学习优化，我们的方法能够更好地挖掘文本与图像之间的协同信息，从而在该类任务中取得优于规则式排序的性能。

综合来看，我们提出的 HiDReader 方法不仅能够有效提升各类模型在文档理解任务中的表现，还具备良好的可迁移性与跨任务适应能力，尤其在复杂结构文档中展现出显著优势。

通过表格的数据可以看出 LayoutLMv3 模型对比 LayoutLMv2 模型，更换阅读顺序的增长较小。由于 LayoutLMv3 模型对图片模态的特征较好，使用了 ViT 对图片模态进行分割和编码，所以图片模态和文字模态的对齐度相对于 LayoutLMv2 模型来说更高。所以重新排列文字模态的顺序的增强没有 LayoutLMv2 显著。同时，在大模型试验中可以显著看出我们的方法对纯文本大模型和多模态大模型都有显著增强，证明根据下游任务进行阅读顺序自主抽取的方法是有效的。所以我们的方法可以在固定语言模型的基础上，通过下游任务的反馈来调整文字的阅读顺序，对语言模型理解文字是有增益的。

(2) 阅读顺序和人类阅读顺序对比

为了更深入地理解 HiDReader 所生成的阅读顺序的行为特征及其与人类阅读习惯之间的关系，我们对模型在实际任务中所产生的阅读路径进行了系统的统计分析与可视化展示。相关结果如图所示。从整体上看，HiDReader 所产生的阅读顺序呈现出与 Z-Order 曲线类似的空间排列特征，即在全局范围内遵循“先左右后上下”的扫

表 3.5　我们在 DocTrack 数据集和 DocVQA 数据集上进行了不同阅读顺序的评估，其中 T 表示 Text，B 表示 Box，I 表示 Image，表中为预训练模型结果

Model	**PREORDER**	**Funsd**			**Seabill**			**InforGraph**	**DocVQA**
		P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	ANLS ↑	ANLS ↑
BERT	EYE	57.75	60.23	58.70	57.63	59.13	58.37	4.01	-
	DEFAULT-OCR	56.69	62.11	60.33	58.99	60.01	59.51	3.82	55.64
	Z-ORDER	**64.09**	**65.28**	**64.66**	63.44	**62.78**	63.11	5.88	57.34
	XYLAYOUT	60.16	60.84	60.19	59.24	60.08	59.65	3.71	53.84
	HiDReader-T	61.13	59.59	60.35	59.58	58.70	59.14	5.59	55.27
	HiDReader-B	61.44	61.80	61.62	60.52	62.45	61.47	6.91	56.83
	HiDReader-B+T	60.17	65.19	62.58	64.15	62.70	63.42	7.02	58.16
	HiDReader-B+T+I	64.06	63.68	63.87	**65.11**	62.33	**63.69**	**7.94**	**58.81**
LayoutLMv2	EYE	82.13	85.11	83.58	77.84	74.14	75.94	16.77	-
	DEFAULT-OCR	86.94	80.95	83.44	**78.56**	73.02	75.69	14.11	78.08
	Z-ORDER	88.00	84.46	**86.06**	78.05	74.77	76.37	21.46	81.42
	XYLAYOUT	84.01	83.12	83.55	75.01	78.41	76.61	15.28	79.27
	HiDReader-T	87.45	83.16	85.25	76.22	77.35	76.78	19.64	77.10
	HiDReader-B	84.63	**85.15**	84.89	76.19	73.09	74.61	17.33	78.99
	HiDReader-B+T	87.92	83.00	85.39	78.37	75.91	77.12	21.26	80.68
	HiDReader-B+T+I	**88.81**	82.93	85.77	78.14	**79.67**	**78.90**	**21.51**	**81.78**
LayoutLMv3	EYE	91.47	91.19	91.33	69.22	65.57	67.35	20.99	-
	DEFAULT-OCR	90.96	92.00	91.48	72.96	66.71	69.70	18.21	83.37
	Z-ORDER	**94.63**	92.85	**93.73**	77.27	68.24	72.47	**24.34**	82.64
	XYLAYOUT	89.09	89.69	89.39	71.05	66.83	68.42	21.02	79.41
	HiDReader-T	90.63	89.92	90.28	74.35	64.61	69.14	18.35	80.19
	HiDReader-B	92.30	91.96	92.13	76.31	66.52	71.08	21.29	81.66
	HiDReader-B+T	92.61	**93.70**	93.15	**77.63**	69.10	**73.12**	21.26	83.45
	HiDReader-B+T+I	93.12	92.40	92.76	75.11	**69.51**	72.20	23.98	**83.87**

表 3.6　我们在 DocTrack 数据集和 DocVQA 数据集上进行了不同阅读顺序的评估，其中 T 表示 Text，B 表示 Box，I 表示 Image，表中为纯文本大模型结果

Model	PREORDER	Funsd			Seabill			InforGraph	DocVQA
		P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	ANLS ↑	ANLS ↑
Qwen1.5-32B	EYE	58.21	22.94	32.91	7.26	15.56	9.90	-	-
	DEFAULT-OCR	60.66	23.29	33.65	7.72	17.27	10.67	58.43	68.34
	Z-ORDER	61.27	**28.01**	**38.44**	11.37	**20.34**	14.59	65.22	72.29
	XYLAYOUT	**68.72**	25.81	37.52	9.49	17.97	12.42	54.31	65.24
	HiDReader-T	55.09	25.03	34.42	8.68	15.77	11.20	58.66	66.30
	HiDReader-B	67.97	24.07	36.55	9.13	16.60	11.78	60.49	70.22
	HiDReader-B+T	67.20	26.44	37.95	10.88	18.21	13.62	63.80	71.60
	HiDReader-B+T+I	59.74	27.55	37.71	**12.27**	19.62	**15.10**	**65.74**	**74.49**
LLaMA3-8B	EYE	74.63	36.02	48.59	39.09	87.10	53.96	26.07	-
	DEFAULT-OCR	53.00	29.44	37.85	36.43	77.86	49.64	18.51	62.48
	Z-ORDER	81.61	**59.43**	**68.77**	37.57	79.89	51.46	**31.67**	61.85
	XYLAYOUT	76.42	42.85	54.91	37.83	78.91	51.14	31.16	56.50
	HiDReader-T	77.56	39.60	52.43	**38.38**	82.59	**52.41**	29.22	60.26
	HiDReader-B	**83.56**	45.86	59.22	35.94	81.72	49.93	28.63	62.85
	HiDReader-B+T	77.72	56.81	65.64	35.66	**84.75**	50.20	30.69	65.23
	HiDReader-B+T+I	73.85	53.71	62.19	37.89	82.47	51.92	31.47	**65.46**

描策略，这种顺序与人类在自然阅读过程中自上而下、从左至右的行为模式具有较高的一致性。

然而，进一步观察其局部阅读行为，我们发现 HiDReader 并不完全依赖于传统的线性阅读顺序。在某些局部区域，模型会选择跳跃式地访问文本信息，即不严格按照“左至右、上至下”的自然顺序进行阅读。这一现象表明，HiDReader 与人类类似，在信息理解过程中能够容忍一定程度上的文字顺序变动，并不依赖文本的绝对语序进行语义建构。换句话说，语言模型具备在非线性顺序下进行有效语义整合的能力，少量文本排列的变化并不会显著影响其整体理解性能。

实验结果如表3.8，在融合了图像模态、文字模态以及位置信息模态的多模态设定下，HiDReader 所生成的阅读路径与人类阅读顺序的相似度最高。特别是在信息图（Infographic）理解任务中，该相似性尤为显著，表明人类在阅读信息图时形成的阅读策略更符合语言模型对于语义组织的理解逻辑。这也进一步说明，在处理具有丰富排版结构的文档时，人类的阅读路径可以为模型提供有效的启发。

表 3.7　我们在 DocTrack 数据集和 DocVQA 数据集上进行了不同阅读顺序的评估，其中 T 表示 Text，B 表示 Box，I 表示 Image，表中为多模态大模型结果

Model	PREORDER	Funsd			Seabill			InforGraph	DocVQA
		P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	ANLS ↑	ANLS ↑
GLM-4V-9B	EYE	43.43	20.54	27.89	47.65	51.56	49.53	64.81	-
	DEFAULT-OCR	41.00	15.91	22.93	48.40	47.29	47.84	62.06	83.19
	Z-ORDER	44.86	22.22	29.72	49.43	**54.34**	51.77	67.20	**84.13**
	XYLAYOUT	43.79	**23.00**	**30.16**	48.84	52.53	50.62	63.77	77.40
	HiDReader-T	43.72	16.70	24.17	48.63	47.46	48.04	64.01	79.04
	HiDReader-B	44.01	20.51	27.98	49.22	47.94	48.57	65.53	80.19
	HiDReader-B+T	42.69	22.39	29.37	51.55	52.85	52.19	67.21	82.61
	HiDReader-B+T+I	**45.81**	22.38	30.08	**51.58**	53.31	**52.43**	**67.83**	83.67
Qwen2-VL-7B	EYE	45.45	11.05	17.78	26.42	6.11	9.92	64.29	-
	DEFAULT-OCR	**56.00**	10.92	18.28	26.92	10.62	15.23	63.81	81.70
	Z-ORDER	51.52	**12.93**	**20.67**	36.90	12.19	18.32	66.30	**85.55**
	XYLAYOUT	54.17	11.04	18.34	**38.65**	9.02	14.63	64.67	82.72
	HiDReader-T	50.38	12.31	19.78	32.34	11.51	16.98	64.13	80.40
	HiDReader-B	51.90	11.81	19.24	35.13	13.27	19.26	65.09	82.78
	HiDReader-T+B	55.43	12.19	19.98	37.70	14.60	21.05	67.02	84.67
	HiDReader-T+B+I	52.60	12.63	20.37	38.32	**14.85**	**21.41**	**67.66**	83.14

此外，我们还进行了多组消融实验，以探究各类模态信息对阅读顺序建模的贡献。实验发现，当仅使用纯文本信息，即缺乏布局和视觉位置信息时，模型在阅读顺序抽取任务中的性能显著下降，所生成的阅读路径与人类阅读顺序的相似度也大幅降低。这一结果表明，仅依赖文字模态的信息不足以决定最优阅读顺序，而视觉布局及位置信息在构建符合人类认知模式的阅读路径中扮演了关键角色。

综上所述，HiDReader 通过引入多模态信息建模机制，能够生成更贴近人类阅读行为的阅读顺序，体现出较强的认知合理性与语义建构能力。

(3) 阅读顺序的迁移能力研究

为验证所提出的阅读顺序抽取模型在不同任务中的适应能力与泛化性能，如表3.9所示，我们在 FUNSD 和 Seabill 两个数据集上进行了下游任务微调，获得了微调后的 **HiDReader** 模型，并将其迁移应用于 SROIE 与 CORD 两个数据集，以评估其跨领域的性能表现。

从实验结果可以观察到，**HiDReader 在未直接参与目标数据集训练的情况下，**

表 3.8　我们分别对三种数据集进行了肯德尔系数和斯皮尔曼系数的统计，其中加粗的为最好的结果，下划线为次优的结果

Modality	**Funsd**		**Seabill**		**InforGraph**		**Overall**	
	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$
BOX	**0.3720**	**0.5046**	0.4346	0.5528	0.7873	0.8468	0.5313	0.6347
TEXT	0.3156	0.4928	0.3914	0.4748	0.4716	0.5272	0.3929	0.4983
TEXT+BOX	0.3299	0.4592	<u>0.5058</u>	<u>0.6235</u>	<u>0.8432</u>	**0.9211**	**0.5596**	<u>0.6679</u>
TEXT+BOX+IMAGE	<u>0.3622</u>	<u>0.5016</u>	**0.6128**	**0.6861**	**0.8872**	<u>0.9142</u>	<u>0.5537</u>	**0.7006**

依然能够在具有相似排版风格的数据集上实现较优的阅读顺序抽取效果。具体而言，SROIE 与 CORD 的版面结构与 FUNSD、Seabill 相似，主要表现为 key-value 信息成对出现，且多以键（key）位于值（value）左侧的形式呈现。这种排版规律使得 HiDReader 能够迁移其在源任务中学到的多模态排序策略，从而在目标任务中表现出良好的阅读顺序理解能力。

值得注意的是，在部分指标上，HiDReader 的表现仍略逊于基于启发式的 Z-Order 方法。这一现象的主要原因在于:HiDReader 并未在 CORD 与 SROIE 数据集上进行针对性的微调训练，导致其对目标域中某些特殊结构或语义模式的适应性略有不足。尽管如此，得益于多模态强化学习策略的引入，HiDReader 能够从下游任务反馈中自动学习并泛化出具有语义合理性的排序策略，展示出良好的跨任务迁移能力。

(4) 可视化结果分析

通过 HiDReader 抽取的阅读顺序可视化结果如图3.4，

为了进一步验证 HiDReader 在文档阅读顺序抽取方面的有效性与合理性，我们对其在典型文档样例中的阅读路径进行了可视化，如图3.4 所示。图中展示了 HiDReader 所生成的阅读顺序的散点图，点的连接顺序即为模型推理得到的阅读路径。

从图3.4中可以观察到，HiDReader 所抽取的阅读顺序整体呈现出一种近似于 Z 字形的结构，即在宏观层面上遵循从左至右、从上至下的通读模式。然而与规则式的 Z-Order 相比，HiDReader 所生成的路径在局部区域表现出明显的聚集性与灵活性。这种局部聚集的特性类似于人类在实际阅读过程中的行为习惯——人类通常会优先聚焦于内容密集的区域，进行局部信息的快速扫读或重点浏览，而非机械性地按照绝对坐标顺序进行遍历。

此外，在信息密度高或排版较为复杂的区域，HiDReader 能够自适应地调整阅读

表 3.9　我们使用预训练后的强化学习模型对 CORD 数据集和 SCORE 数据集进行了迁移学习

Model	**PREORDER**	**CORD**			**SROIE**		
		P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑
BERT	DEFAULT-OCR	92.26	93.03	92.64	84.97	**89.91**	86.99
	XYLAYOUT	79.47	77.02	78.22	81.87	85.49	83.64
	Z-ORDER	85.51	85.14	85.33	84.50	88.35	86.38
	HiDReader	**92.84**	**93.21**	**93.02**	**85.01**	89.51	**87.20**
LayoutLMv2	DEFAULT-OCR	92.48	**90.39**	91.37	87.72	**91.57**	88.54
	XYLAYOUT	86.87	84.20	85.51	80.01	87.89	83.78
	Z-ORDER	87.21	88.36	87.78	85.29	91.35	88.20
	HiDReader	**92.97**	90.34	**91.63**	**88.81**	90.63	**89.71**
LayoutLMv3	DEFAULT-OCR	90.42	90.83	90.62	87.66	85.95	86.80
	XYLAYOUT	89.17	86.36	87.74	82.11	83.18	82.64
	Z-ORDER	90.12	90.78	90.44	88.17	**86.82**	**87.49**
	HiDReader	**91.19**	**91.15**	**91.17**	**88.70**	85.92	87.29
Qwen1.5-32B	DEFAULT-OCR	50.45	36.52	42.37	71.06	41.59	52.47
	XYLAYOUT	38.55	31.37	34.59	45.76	35.94	40.26
	Z-ORDER	**56.02**	40.63	47.10	70.42	**46.11**	**55.73**
	HiDReader	54.70	**45.40**	**49.62**	**73.46**	41.92	53.38
LLaMa3-8B	DEFAULT-OCR	8.59	15.41	11.03	51.39	63.44	56.79
	XYLAYOUT	7.71	11.63	9.27	52.21	72.10	60.57
	Z-ORDER	**17.87**	**13.84**	**15.60**	50.98	**77.07**	61.36
	HiDReader	11.47	13.56	12.43	**56.32**	73.41	**63.74**
GLM-4V-9B	DEFAULT-OCR	55.30	11.82	19.48	18.49	**38.82**	25.05
	XYLAYOUT	46.43	9.28	15.47	**21.34**	31.53	25.45
	Z-ORDER	**58.04**	14.62	23.36	21.31	38.52	**27.44**
	HiDReader	57.17	**15.33**	**24.18**	19.19	37.99	25.58
Qwen2-VL-7B	DEFAULT-OCR	42.27	8.56	14.23	41.99	14.42	21.47
	XYLAYOUT	35.29	9.16	14.55	48.21	17.43	25.60
	Z-ORDER	43.63	**12.37**	**19.27**	47.74	**18.71**	**26.88**
	HiDReader	**46.52**	10.56	17.18	**48.38**	17.51	25.71

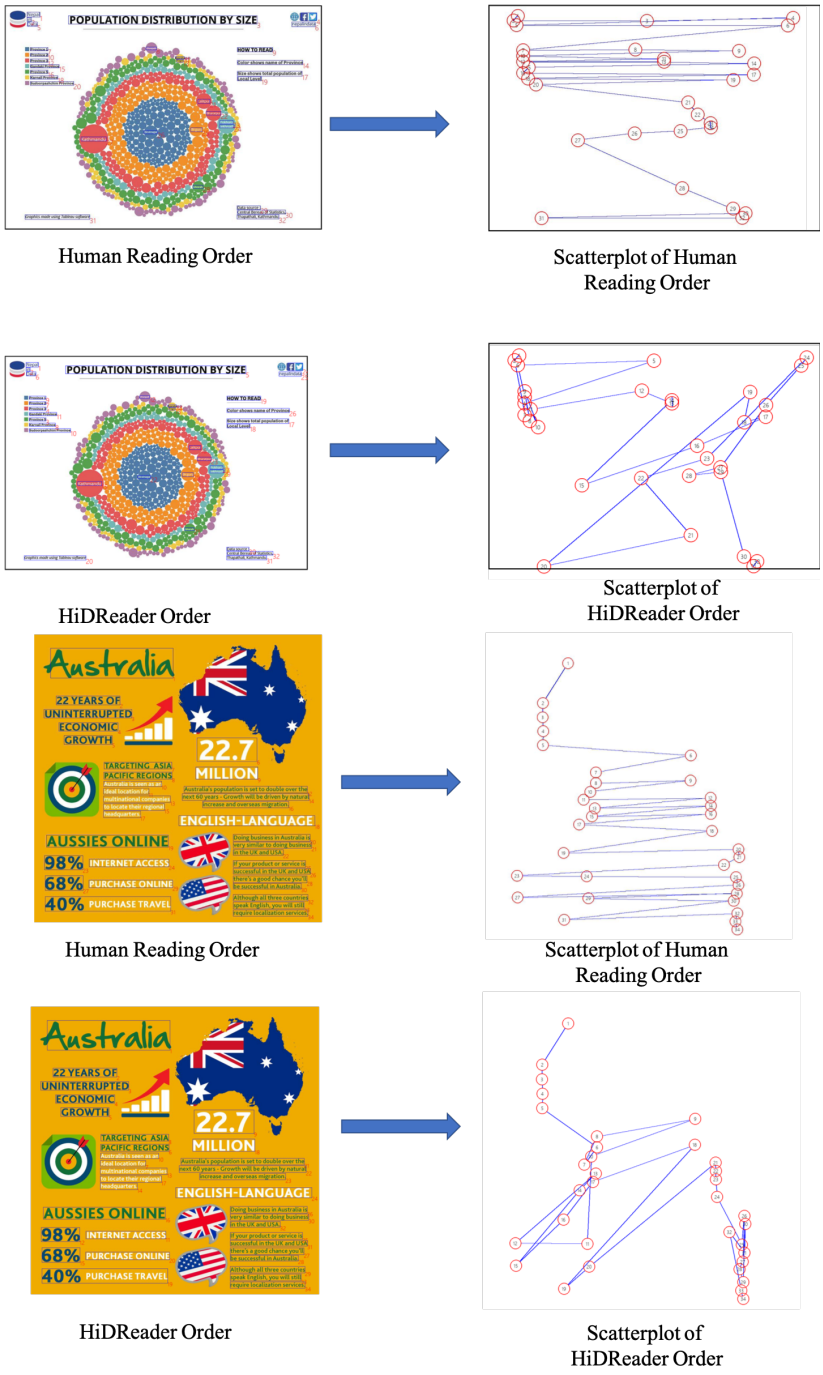


图 3.4 我们的阅读顺序和人类阅读顺序的对比

路径，以更高效地捕捉局部语义关系。这种灵活的局部跳转能力不仅提高了模型的信息整合效率，也增强了其对真实文档结构的理解能力。因此，我们认为 HiDReader 所抽取的阅读顺序在保持整体逻辑通顺的同时，也兼顾了人类阅读行为中的局部策略性，是一种更为合理和智能的阅读路径建模方式。

(5) 消融实验

为进一步探究不同模态信息对 HiDReader 模型在阅读顺序抽取任务中的贡献，

我们设计并实施了系统性的**消融实验**。本实验围绕四种输入模态配置展开: 仅使用文本信息(Text-Only)、仅使用坐标信息(BBox-Only)、结合文本与坐标信息(Text+BBox),以及融合文本、坐标与图像信息(Text+BBox+Image)。相关实验结果如表 3.5-3.7所示。

从表中可见,**坐标信息在阅读顺序抽取中起到了关键作用**,单独使用坐标信息(BBox-Only)即可取得较为优越的排序效果。这与人类的阅读机制相似: 即使不理解文字语义,人类依然可以凭借排版布局判断基本的阅读顺序,说明布局特征具有天然的排序指示性。

当结合文本与坐标信息(Text+BBox)输入时,模型的性能进一步提升。这一结果表明,**文字与位置信息的融合有助于捕捉文档中的语义与结构信息**,特别是在诸如键值对(key-value)结构识别等任务中,显著增强了模型的理解与排序能力。图 3.5 所示为大模型在仅使用文字和坐标信息输入的情况下生成的文档布局示意图,展示了多模态协同下模型对结构的良好还原能力。

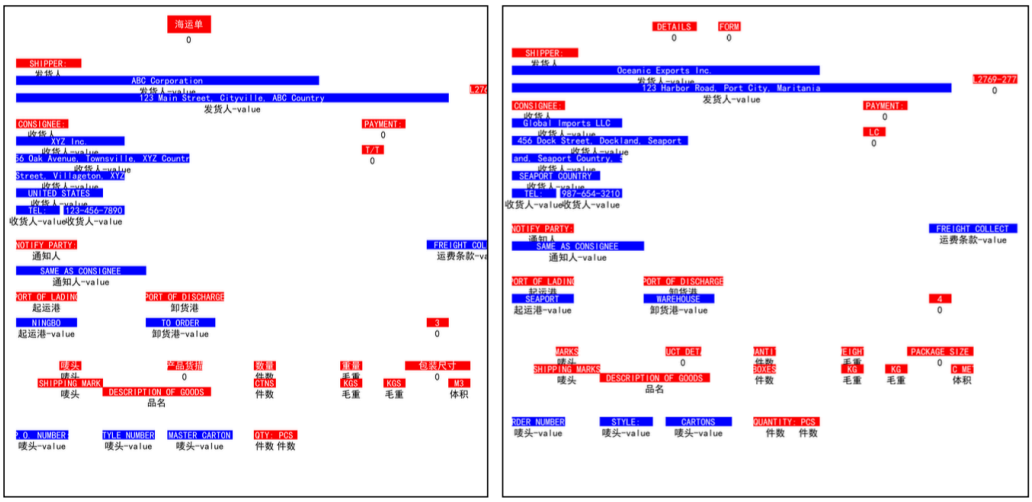


图 3.5　大模型基于文字与坐标信息生成的布局结构示意

在进一步引入图像信息(Text+BBox+Image)后,虽然模型理论上可以从中获得更丰富的上下文与视觉线索,但我们观察到性能提升并不总是稳定。我们推测,这种现象主要源于三模态融合过程中存在的**对齐误差与信息噪声**,特别是在强化学习框架下,由于训练过程中的策略波动性,引入更多模态可能加剧策略空间的不确定性,反而导致性能波动或下降。

综上所述,消融实验验证了不同模态对阅读顺序抽取任务的相对贡献,其中坐

标信息的作用尤为突出，而文字与坐标的融合可实现最佳的任务适应性。多模态融合虽具备潜在优势，但在实际训练中仍需谨慎设计其对齐与融合策略，以避免噪声干扰模型学习过程。

3.5.5 本章总结

本章提出了一种基于强化学习的自主阅读顺序探究方法，该方法通过将阅读路径优化建模为强化学习问题，使智能体能够在无监督的情境下，借助下游任务的反馈信号和语义相似度约束，自主学习并生成最优的阅读顺序。该策略跳脱了传统预设规则所带来的限制，使得阅读顺序可以动态适应不同类型任务的需求，增强了模型在实际应用中的灵活性与泛化能力。

在实验部分，我们验证了所提方法在多种典型自然语言处理任务中的有效性，包括文档分类、信息抽取与图文理解等。实验结果表明，基于该强化学习策略所学习到的阅读顺序不仅在性能指标上优于传统的固定顺序策略，而且在阅读路径的可视化结果中，其阅读行为更贴近人类自然的阅读方式，呈现出“全局有序、局部灵活”的阅读特征。这种阅读模式在宏观层面上展现出类似 Z 字形的结构，但在微观层面上允许一定的跳跃式阅读，即局部区域的非线性访问，显示出语言模型具备灵活的信息组织与理解能力。

此外，通过对奖励函数的精心设计，我们进一步提升了模型对下游任务的适应性，保证智能体所学得的阅读顺序能够有效提升具体任务的性能。在跨任务迁移学习实验中，HiDReader 亦表现出良好的泛化能力，表明该方法具有较强的迁移潜力，能够适用于结构不同但语义目标相似的多种文档场景。

所以综上，本章所提出的强化学习阅读顺序建模方法不仅提升了模型在自然语言理解任务中的表现，也从认知合理性与行为策略上推动了机器阅读路径优化的研究，为实现更高效、更人性化的机器阅读系统提供了重要思路和实践基础。

第四章　基于智能体人类对齐的多页文档阅读顺序生成方法

4.1　研究动机

在现实生活中，文档类型呈现出高度多样性，包括发票、海报、说明文档、报告书、合同等结构化程度不一、页面数不等的文档形式。面对这些异构文档，传统的阅读顺序抽取方法通常依赖于大量手工标注的单页文档数据，主要聚焦于单页或者结构较为简单的短文档场景，难以扩展到真实环境中常见的多页、篇幅较长的文档处理任务,并且随着文档长度的增加，具有截断可能的段落数量就越多，不合理的阅读顺序会对多页长文档理解任务造成极大的限制。因此，研究一种能够**跨文档长度**、具有**通用能力**的阅读顺序抽取模型，成为推动文档理解任务进一步发展的关键方向。

通用的阅读顺序抽取模型[73-75]不仅能够提升文档问答、关系抽取、关键信息定位等下游任务的性能，还能够为大模型提供更合理的输入结构，从而提高模型的感知和推理能力。然而，多页长文档往往存在结构复杂、阅读路径分散、视觉/文本层级丰富等特点，导致基于强化学习的自主探索在此类场景中难以稳定收敛，泛化能力不足。同时，学术界当前尚缺乏面向多页长文档的、具备人类阅读行为指导的系统性阅读顺序标注数据集，使得构建能够泛化于长文档阅读任务的模型面临较高挑战。

近年来，以大语言模型和多模态大模型为代表的大模型范式在多个自然语言处理任务中展现出强大的知识表达与任务迁移能力。通过轻量级的提示构建或少量样本的任务微调，便可以赋予模型较强的任务执行能力。此外，大模型还呈现出随参数规模和数据规模增长而出现的新能力，为复杂任务处理提供了新的可能性。因此，将大模型应用于阅读顺序的理解与抽取，尤其是长文档中的通用顺序建模，具有高度的前沿性与实用价值。

尽管如此，当前主流的大模型微调方法在面对超长文档阅读顺序建模任务时仍存在明显瓶颈。一方面，序列长度限制与计算资源开销制约了模型对整篇长文档的全局感知能力；另一方面，强化学习等优化范式难以直接迁移至高维、稀疏反馈的长文档排序场景。在此背景下，引入人类眼动顺序标注等具备认知指导价值的监督

图 4.1　M6Doc 多页文档示例

信号，成为提升大模型在阅读顺序抽取任务中有效性的重要路径。然而，当前缺乏大规模、系统化的人类眼动顺序对齐数据集，也使得该研究方向的发展受到掣肘。

基于以上问题分析，本文聚焦于“通用长文档阅读顺序抽取大模型”的构建与研究，致力于填补当前领域在长文档阅读顺序建模上的空白。具体而言，我们设计并构建了三类不同长度文档的阅读顺序数据集，分别涵盖单页文档、中长文档和超长文档，均以人类眼动顺序或主观阅读路径为参考进行精细化标注，力图还原真实阅读过程中的注意力转移轨迹。在此基础上，我们进一步提出了一种融合人类阅读顺序对齐策略的通用大模型训练方法，构建了一个可用于多种文档类型与长度场景的阅读顺序抽取模型。实验结果表明，该模型在文档问答与命名实体识别等典型下游任务中均表现出显著优于传统方法的性能，验证了其在长文档阅读理解中的有效性与通用性。

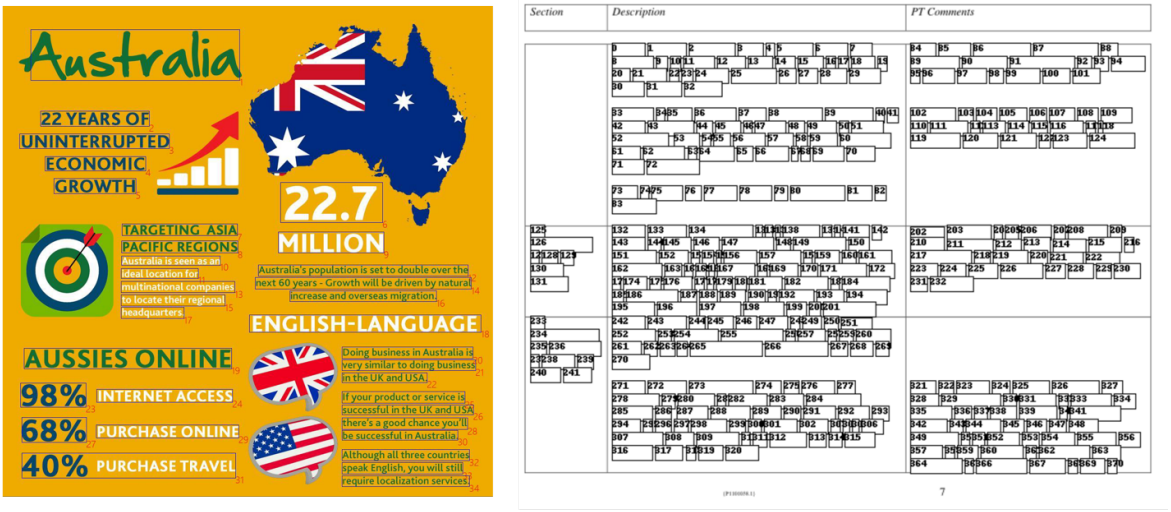
4.2 方法对比

4.2.1 文档阅读顺序建模相关研究

随着多模态文档理解任务的发展，文档中的阅读顺序建模（Reading Order Modeling）逐渐成为影响模型性能的关键因素之一。合理的阅读顺序不仅能够提升文档结构重建与信息抽取的准确性，还可以显著增强下游如问答、摘要等任务中的语义一致性与推理能力。因此，构建具备高质量阅读顺序标注的数据集成为当前研究的重点方向之一。

近年来，研究者相继提出了若干面向文档阅读顺序的基准数据集，其中具有代表性的是 ReadingBank 和 DocTrack 两个数据集。ReadingBank 由微软亚洲研究院构建，面向结构化 Word 文档（包括.doc 和.docx 格式），通过解析 XML 源文件，提取出具有人类书写顺序的正文内容，并结合其在页面中的空间位置信息进行标注。该数据集涵盖多栏、表格等复杂文档结构，旨在提升模型在复杂排版环境下的信息提取与结构感知能力。

而 DocTrack 数据集则由王等人提出，致力于构建更贴近人类阅读行为的阅读顺序标注机制。其核心方法是通过大量人类参与标注，融合主观阅读路径与任务导向的关注顺序，从而训练模型更好地模拟人类在真实环境中的阅读策略。DocTrack 强调多主体对同一文档的阅读多样性，在对齐过程中引入人类行为的统计特征，在一



DocTrack　　Readingbank

图 4.2　DocTrack 数据集和 ReadingBank 数据集的可视化

定程度上弥合了模型与真实阅读习惯之间的差距。

尽管上述数据集在推动文档阅读顺序建模研究方面取得了积极进展，但仍存在显著不足：这些工作多聚焦于单页文档的阅读顺序建模，尚未充分考虑跨页场景下信息组织与用户关注之间的动态变化。

为填补上述空白，我们构建了一个面向多页结构化文档的人类眼动对齐阅读顺序数据集。该数据集融合真实眼动轨迹与文档排版信息，能够更准确地反映用户在长文档中的自然阅读路径，为后续构建更贴近人类认知机制的文档理解系统提供了关键支撑。同时，本数据集有助于推动强化学习、模态对齐等方法在阅读顺序建模中的落地与发展，为长文档理解开辟新的研究方向。

4.2.2　阅读顺序提取方法对比

传统的阅读顺序抽取方法可以分为规则型，例如 Z 字顺序和 XyLayout 阅读顺序抽取，这两种方式是基于规则的阅读顺序抽取方法，具有抽取速度快，但是抽取模式过于僵化的特点。另一种阅读顺序抽取方式是基于模型的抽取方式，例如 LayoutReader 使用大量标注的 xml 数据集进行模型拟合，TPP 模型利用了图神经网络的能力，对阅读顺序进行拟合和模拟。我们的方法主要应用了大模型的能力，通过构建 prompt 来对阅读顺序进行提取[76-78]。我们通过构建模板化的任务描述对阅读顺序抽取任务进行了规范和约束。

表 4.1 大模型多模态任务的提示构建方案

（1）提示构建：本小节介绍如何为大模型构建合理的提示信息，包括指令构建与查询指令设计两部分。

任务描述（I_t）：大模型需要清晰的任务描述来调用自身庞大的知识库。以问答任务（QA）为例，一个合理的任务提示如下：

你的任务是根据文本回答问题，对于给定的问题 xxx，请根据给定的文本信息 xxx，充分理解文本的语义，作出回答。

该提示信息可以引导模型理解语义并准确作答。

（2）标签映射（I_l）：问答任务中需预测正确答案，提示中应明确模型的预测空间。标签映射旨在将语义相近的内容统一表示，例如“红色”表示“图像中物体的颜色”。

标签映射形式如下：

$$I_l = (Y_1', Y_1), (Y_2', Y_2), ..., (Y_n', Y_n)$$

（3）上下文范例：基于上下文学习，在零样本基础上加入范例可增强模型理解。以视觉问答为例，上下文范例包括：

- 图像描述：图像内容 I_c 及其关键区域位置信息 B_c
- 问题：如“图像中物体是什么颜色?”
- 正确答案：如“红色”

上下文范例可形式化为：

$$ContextDemonstrations = (I_c, B_c, Q_c, Y_c)$$

（4）约束条件：为防止模型生成无意义或幻觉内容，在提示中需加入约束信息，限制其输出范围，提高语义合理性。

（5）格式化输出：为便于结果解析与评估，需规范模型输出格式。例如在 QA 任务中，要求输出格式为 JSON：

{ "answer" : " 红色" }

此结构化格式便于后处理。

（6）查询提示（Q）：查询提示由问题对象（如图像区域）与具体问题组成，用于引导模型关注特定内容。形式如下：

$Q = \{QuestionObject, Question\}$

例如，问题对象为图像中某个区域，具体问题为“该区域物体的颜色是什么?”

(7) 推理过程：构建好提示后，模型根据图像 I 和问题 Q，结合指令 I_{inst} 与查询提示 Q_{query}，生成最终预测结果 A:

$$A = LLM(I_{inst}, Q_{query}, I, Q)$$

输出结果遵循标准格式，便于下游应用如问答系统、图像检索等场景中的使用。该方法无需微调即可实现高效推理，展示了模型强大的泛化能力与多模态理解能力。

4.3 方法介绍

图展示了我们的方法流程，通过多个步骤来对齐人类的阅读顺序，并且在下游任务中对输入到 LLMs 中的阅读顺序进行预排序增强 LLMs 对长文档的理解能力进行提升，为了收集人类的阅读顺序，我们使用了见数 (Credamo)[①]平台进行了阅读顺序的收集。受试者阅读的 500 篇文档来自于 LongDocURL,M6doc 和 DocVQA 数据集，涵盖了超长文档、中等长度文档和单页文档多种数据集。这些文档涵盖了多种格式，包括简单的表格，布局复杂的海报，多栏的学术论文等。这些数据集具备不同的布局信息，可以有效的帮助大模型在微调时获取不同的布局信息。

4.3.1 人类阅读顺序收集

(1) 人类眼动顺序收集

1) 研究伦理与参与者权益保护。

在研究开始之前，我们对所有参与标注的同学详细的说明了实验的目的、流程、潜在的风险以及他们的权利，确保每一位参与者都充分理解实验中的相关内容，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。

2) 眼动设计实验和数据采集

为了开展眼动实验，我们将文档数据集划分为了三种类型，第一种是长度超过 50 页的长文档，第二种是长度在 2-20 页的中等长度文档，第三种是单页的文档。并且招募了 10 位以上的在校研究生进行实验参与者，实验参与者都具有理解英文文本的能力。实验采用见数 (Credamo) 平台进行眼动数据采集和标注。实验过程中，受试

① https://www.credamo.com//

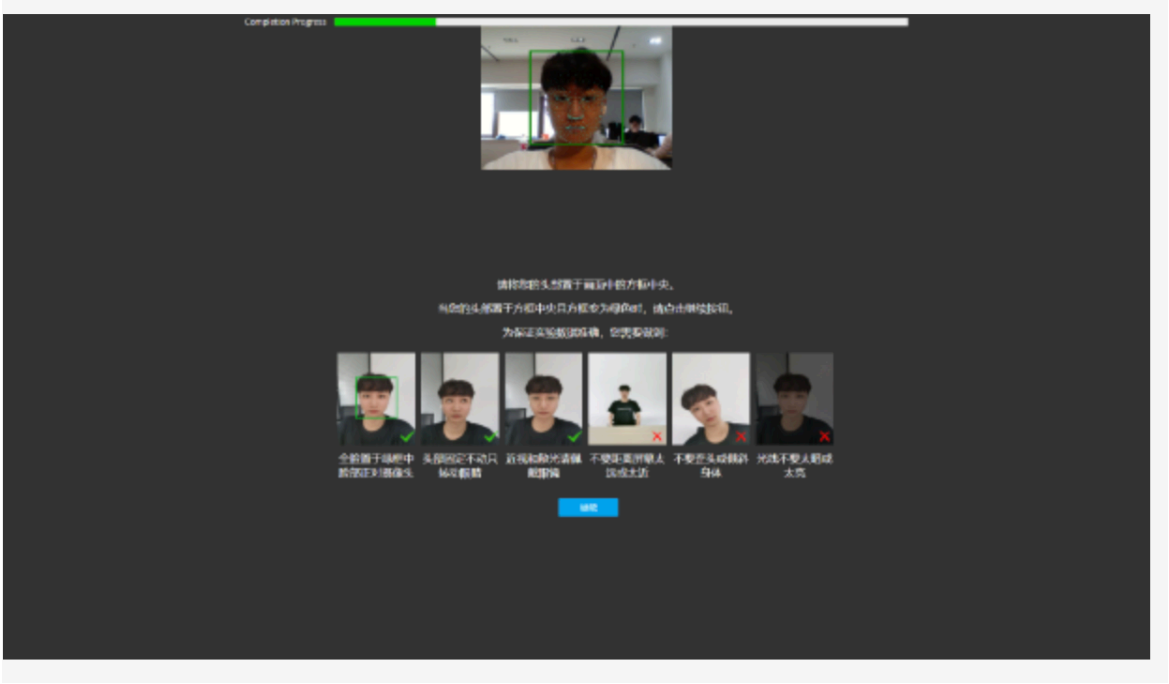


图 4.3　见数平台校准标准

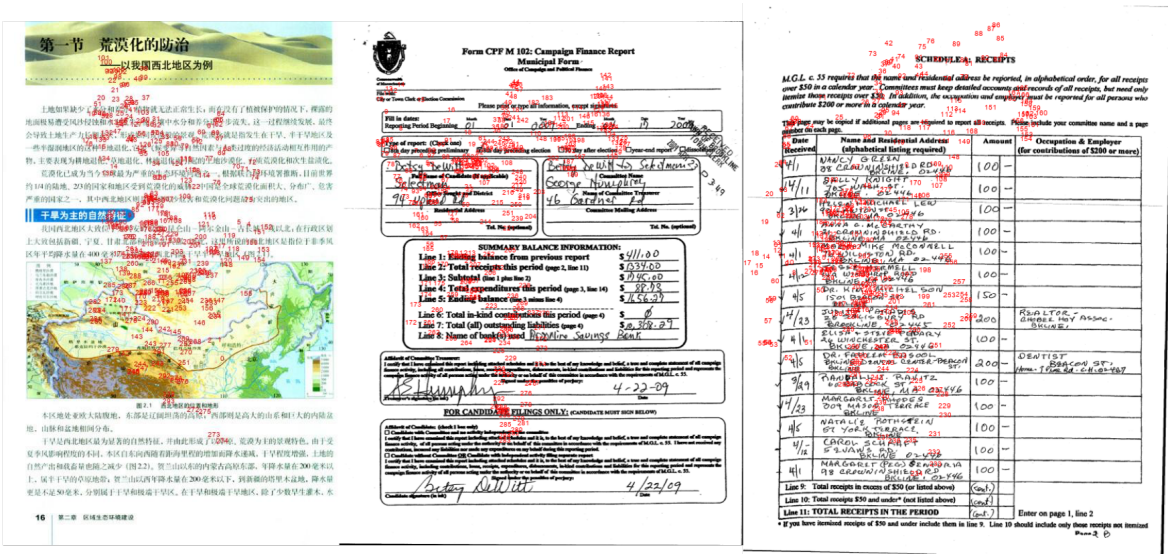


图 4.4　数据集部分人类标注数据展示

者首先需要通过眼动锚定，来确认自身的眼动是否可以被摄像头成功记录，如图片所示，受试者需要保持自身的位置尽可能不发生变化，视线始终锚定在摄像头的误差范围以内，从而进行眼动标注，这样的结果才会被接纳。受试者通过电脑屏幕中的提示信息进行阅读行为，并且通过摄像头记录下了眼动的数据，具体包括阅读顺序，阅读坐标、阅读时间等数据，这些数据为后续的文档理解任务提供了大量的参考数据。如图所示，我们展示了数据集的部分标注数据4.4。

(2) 实验数据预处理

由于眼动仪存在着高频率的采样过程，原始记录的数据和实际的注视位置存在着一定的偏差，此外受到余光的影响，部分的注视点可能存在误差。为了提高数据的准确性，我们对数据做了以下的平滑:

(1) 对注视点缺失的情况，我们采取如果落在 OCR 边缘或在一定的欧式距离内，则我们认为该注视点在对应的文本框附近。则采纳该文本框作为我们的当前阅读点。

(2) 对于重复阅读的词语，由于人类在阅读过程中容易重复阅读已经阅读过的词语，所以我们将重复阅读过的词语只取第一次阅读的顺序作为阅读顺序。

(3) 如果注视缺失点不属于前两种情况，我们将前两个注视点之间的数据作为当前阅读顺序，从而填补上差值。

(3) 标注一致性检验

在实验的最后阶段，我们将从多名参与者中选择不同的参与者对同一份文档进行阅读顺序再次标注，从而获得最终的阅读顺序。对于同一份文档的标注信息，我们将采纳多名标注者的平均标注结果作为最终的标注结果。从而大大提升阅读顺序标注的鲁棒性和适应性。并且所有参与标注的受试者均存在一定的科学基础，所以可以为后续的高质量标注提供规范性和可靠性。通过上述的实验收集方式，我们可以得到一批具有较高质量的眼动数据，为后续实验的进行提供了强有力的支持。数据集收集的信息如表4.2

表 4.2　数据集信息统计

	LongDocURL	**M6Doc**	**DocVQA**
标注数量	100	100	300
平均页数	74	3	1
语言	English	Chinese, English	English
实体类别	3	74	5
文档类型	PDF, Scanned, Photographed	PDF, Scanned, Photographed	PDF

4.3.2　基于大模型的阅读顺序抽取方法

为了充分训练我们的阅读顺序大模型，我们需要对阅读顺序进行多维度的标注，我们的阅读顺序生成大模型分成了三个部分，首先是阅读顺序的人类标注让大模型具有黄金标签，其次是大模型的 GRPO 微调，让大模型可以根据人类的阅读顺序通过奖励函数的激励学得阅读顺序。最后再根据我们获得的阅读顺序大模型对评估的

文本进行预排序。

(1) 大模型基座对齐人类

在大模型的排序过程中，如图所示，我们通过将文档内容输入到大模型中，通过给定一定的 prompt，如下：

“如图所示，我们需要将上述文档内容按照人类的阅读顺序进行阅读，请直接告诉我文档中的文字阅读顺序，在 <answer></answer> 中输出，仅给出文字内容不需要额外的分析。”

通过这样的 prompt 激发大模型转化的能力，可以让大模型解析图片并且根据图片中的内容输出对应的文档文字内容。[79-82]

(2) 奖励函数设计

根据上述收集数据的方法，我们获得了一批高质量的人类阅读顺序标签数据，我们需要让大模型解析的内容对齐人类的阅读习惯，所以我们根据提取出的阅读顺序对大模型的输出结果进行规范化对齐。根据和人类的阅读的相似性进行打分操作，例如使用向量相似度进行打分，所以奖励函数的第一部分就是根据同人类阅读顺序的相似性进行奖励。[37,83-84]

$$R_s = sim(T_h, T_l) \tag{4.1}$$

这里的 T_h 代表人类的阅读顺序，T_l 代表大模型输出的阅读顺序。通过输出文本的相似度，我们就可以给定大模型输出好与坏的标准，有效的让大模型输出的阅读顺序和人类的进行对齐。

第二部分的奖励是规范化输出的奖励，由于大模型的输出结果容易发生幻觉，即大模型根据输出的内容进行联想，从而导致输出的结果发生偏差，不够精准，所以我们增加规范化输出奖励，即只有输出的结果在标签 <answer></answer> 中，我们才给予规范化输出奖励，并且随着大模型输出的内容增加，奖励相应的会减少。即约束大模型的输出结果，让它不要输出与本次 prompt 无关的内容。

$$R_r = \begin{cases} 0.3 - T/T_m, \\ 0 - T/T_m \end{cases} \tag{4.2}$$

这里的 T 表示当前文本长度,T_m 表示最长文本的长度，通过这样的输出结果规范化，我们可以有效的规避大模型过长的输出结果导致可能存在的幻觉问题。并且可以让

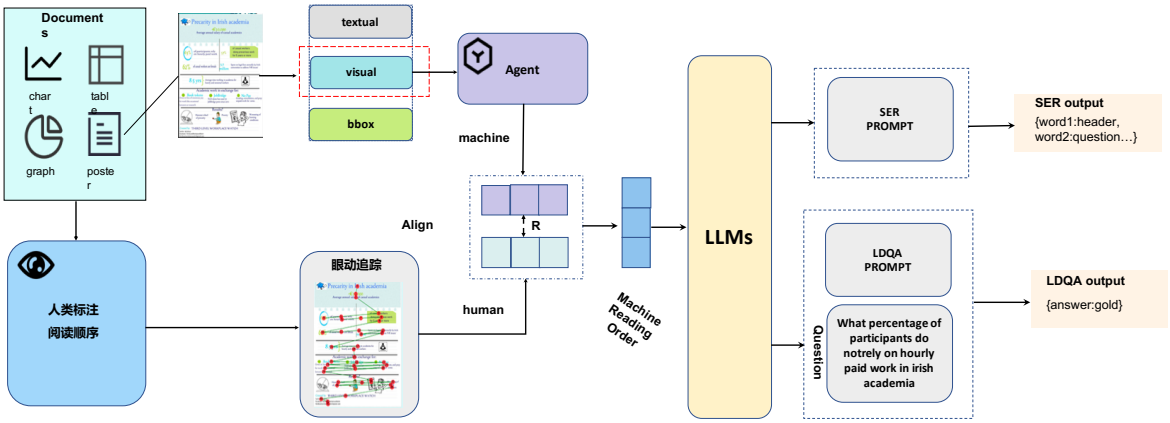


图 4.5 我们的整体处理流程

大模型能够按照规范化的结果进行输出，方便我们在下游任务中对产生的阅读顺序进行提取和解析。

4.3.3 基于大模型的文档理解方法

我们将上述得到的阅读顺序输入到大模型中进行评估，我们的实验包含三个阶段，分别是数据预处理，阅读顺序构建，下游任务评估。如图4.5所示 我们首先将文档内容输入到大模型中，经过大模型处理后，得到包括阅读顺序的输出中间结果，再通过人类的阅读顺序标签对大模型输出的结果进行对齐，通过 GRPO 的方式训练得到靠近人类阅读习惯的大模型智能体，最后再将重新排序后的阅读顺序输入到下游评估的模型 LLM 中，通过精心设计的提示信息明确告诉这些模型要执行的任务。在 LLM 中，我们仅给出这些文本信息 t_i 和文本信息对应的坐标信息 x_i，y_i，这些信息用来提供文本信息和辅助大模型理解的坐标信息。这种根据大模型能力区分的输入内容，充分的利用了 LLMs 在文本理解上的优势。让模型可以更加高效的处理文档任务，具体的公式如下:

$$L_i = LLMs(t_i; x_i, y_i; Prompt_1) \tag{4.3}$$

4.4 任务定义

为了评估我们阅读顺序提取的质量，我们需要在下游任务中进行分析，经过我们重新排序之后的文本需要在下游任务中获得更强的指标表现。所以，我们为了充分衡量经过重排序之后的阅读顺序是否增强了大模型的理解能力，我们进行了两种任务进行阅读顺序的衡量，分别是长文档信息问答任务和命名实体识别任务。长文

档信息问答任务可以衡量 LLMs 对整体文档的理解能力，命名实体识别任务可以衡量 LLMs 对局部信息的理解能力。

4.4.1 长文档信息问答任务

随着多页文档在实际应用中的广泛存在，长文档信息问答任务（Long Document Question Answering, LDQA）成为文档智能理解领域的重要研究方向。该任务旨在基于超长文档中的内容，对指定问题进行自动化回答。不同于传统的短文档问答任务，LDQA 任务需要模型具备跨页整合信息、对抗阅读顺序错乱、以及处理内容截断等复杂挑战。

在长文档场景中，信息往往分布于多个页面，且文本块之间缺乏明确的线性结构。由于模型的最大输入长度限制，单次无法处理整个文档的全部信息，因此需通过高效的文本提取与排序策略，将文档内容重构为合理的输入序列，辅助模型理解上下文逻辑关系。相比之下，短文档问答任务通常基于单页文档，内容结构紧凑，信息关联明确，不易出现阅读顺序错乱的问题。

图4.5中给出了一个典型的 LDQA 任务示意。整个流程可分为三个阶段：（1）**文档文字提取**：使用 OCR 引擎对多页 PDF 或图像文档进行识别，获取所有页面的文字内容及其位置信息；（2）**阅读顺序重建**：采用排序模型或规则方法对提取的文本块进行排序，使其更接近人类的自然阅读顺序；（3）**大模型问答推理**：将排序后的文本信息作为上下文输入，结合任务问题 Q 及提示词（prompt），通过大语言模型生成答案 $\hat{A}$。

在该任务中，阅读顺序的准确性对问答效果具有显著影响。如果排序结果出现段落错位或上下文割裂，模型容易生成不相关甚至错误的答案。因此，排序质量成为影响 LDQA 性能的关键因素。**评价指标**方面，本文采用 ANLS 作为主要评估标准，并且使用了 F1 值作为回答质量的判断。ANLS 和 F1 综合考虑了答案与参考答案之间的语义和表述差异，具体定义与实现细节见第三章。

4.4.2 基于大模型的迁移学习

我们训练对齐得到的阅读顺序抽取模型应当可以适用于多种场景，所以我们需要对大模型的阅读顺序抽取结果进行迁移验证，我们采用未标注的 M6Doc 数据集中的其他类型文档，例如我们训练时使用的是笔记、试卷等类型，我们验证时使用

M6Doc 数据集中的书籍、论文等类型的数据进行迁移学习从而可以验证我们的阅读顺序是否具备通用理解能力。并且我们使用 SER 任务进行迁移评估。并且我们将 SER 任务在这里分为粗粒度的 SER 和细粒度的 SER，由于 M6Doc 数据集中只有细粒度的标注信息，所以我们将 M6Doc 中的实体根据文档结构划分为了三种类型，分别是标题、正文和其他进行粗粒度的实体判别。

4.4.3 顺序相似度评估指标

在顺序相似性评估任务中，斯皮尔曼相关系数和肯德尔系数是常用的统计方法，主要用来统计两个分布的单调关系。这两种统计方法适合于处理序列顺序的数据，用于衡量序列之间的关系，无需让数据符合正态分布，主要应用场景是 **1) 评估序列一致性:** 计算斯皮尔曼系数和肯德尔系数可以衡量两个时间序列的相似度，阅读顺序是一个符合时间序列分布的数据。**2) 评估非线性关系的相似度：** 这两种系数对非线性关系的敏感度较高，适用于分析相关数据。**3) 处理秩次数据：** 当数据类似于排名数据时，使用这两个更为合适。**4) 鲁棒性:** 当数据为异常数据时，这两种系数的敏感度更低。

肯德尔系数较多的被用来衡量两个序列之间的一致性，需要在衡量全局局部排序一致性的时候使用肯德尔系数，斯皮尔曼系数用来衡量两个序列之间的单调关系，可以捕捉全局的趋势。具体的计算方式见第三章内容。

4.5 实验和分析

4.5.1 数据集

实验数据集我们使用了三个公开的数据集，LongDocURL，M6Doc,DocVQA 从中挑选出了部分的数据进行阅读顺序收集。这些数据集的统计信息如表4.2

LongDocURL：从这个数据集中挑选了 100 个超长 PDF 文件，每个 PDF 文件都有一个问题和对应的答案。

DocVQA:DocVQA 是一个专门为视觉访问的数据集，通过图片中的文字布局来进行问题回答，这个数据集可以充分的利用文档中的复杂图片和布局信息对预先准备好的问题-答案对进行回答。我们从该数据集中选择了 300 张图片进行了人类眼动标注。

M6Doc 数据集: 这个数据集包含 9,080 张现代文档图像，这些图像被精心划分为七个不同的子集：科学文章、教科书、试卷、杂志、报纸、笔记和书籍。这些文档不仅覆盖了 PDF、拍摄文档和扫描文档三种常见格式，还提供了 237,116 个详尽的标注实例。在深入分析不同文档类型之间的标注标签的共性与特性、标签的出现频率以及独立页面的识别等关键因素后，我们最终使用 Qwen2.5VL-72B 模型对这个数据集中的文档生成了 100 个 QA 对。

4.5.2 实验环境和细节

为了微调 Qwen2VL-3B 模型，我们使用了 3 张 A100, 其中 2 张用来加载模型，1 张用来做 vllm 推理，我们使用 ms-swift 框架进行微调和推理[85]，为了充分考量我们的通用阅读顺序对 LLMs 的影响，我们在 LLama3-8B,GPT4o 等纯文本大模型上进行了多种阅读顺序的对比。实验的环境如表4.3 我们使用的大模型具体参数如表4.4

表 4.3　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	A100
Python 版本	1.8.1
Pytorch 版本	3.8.10
CUDA 版本	11.4
transformer 版本	4.5.0

表 4.4　模型信息统计

模型	**参数量**	**是否开源**	**提出时间**
LLaMA 3-8B	8 billion	是	2023
Qwen2-7B	7 billion	是	2023
GPT-4o	未知	否	2023
GLM4-9B	9 billion	是	2023

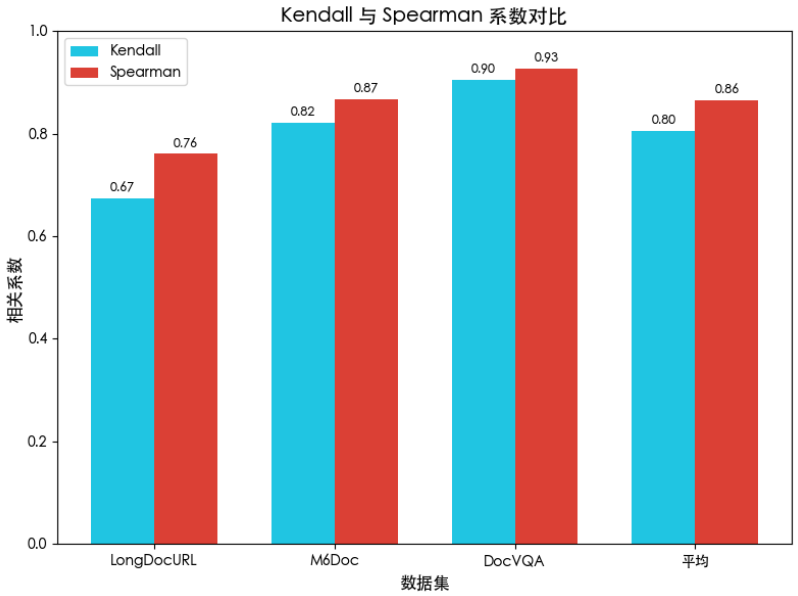


图 4.6 肯德尔系数和斯皮尔曼系数的对比

4.5.3 顺序相似性评估结果

我们在四种不同的大模型上在三个不同的数据集上进行了序列相关性的评估结果4.6，研究的结论如下，通过人类阅读顺序的激励，通过 GRPO 微调我们的大模型可以掌握多种长度文档的阅读顺序，我们的阅读顺序和人类的阅读顺序具有较高的相似度。即大模型可以通过人类的阅读标签获得和人类类似的表现结果。这证明了大模型具备类似人类的阅读方式，可以有效的学到人类的阅读习惯。并且通过观察，结构化强的文档可以更有效的获得和人类更相似的阅读顺序。结构化弱一点的文档的阅读顺序和人类的相差会更大一些。从图中可以看到 DocVQA 的相似度最高，由于这个数据集都是单页的文档，并且布局信息较为复杂，所以可以获得更好的显示效果。其中 LongDocURL 的相似度是最低的，因为这个数据集的数据页数较多，涵盖了多种文字的分布方式，所以大模型获得的阅读顺序和人类的相似度最低，最后是 M6Doc 这个数据集比较结构化数据比较密集，并且长度适中，所以大模型的阅读顺序和人类的阅读顺序相似度比较中等。

4.5.4 文档问答任务结果分析

我们首先在默认阅读顺序，Z 字顺序，XYLayout 顺序，人类阅读顺序 (Human) 和大模型模拟顺序 (Ours) 这五种阅读顺序中进行了实验，为了充分验证不同阅读顺序带来的 QA 性能提升，我们采用了多种大模型进行了验证，可以从结果中看出，人类

的阅读顺序确实对长文档的理解任务有促进作用，在 LongDocURL 中由于大模型无法输入过长的文本，所以我们使用了 RAG 技术进行处理，即将文本转化为 MarkDown 格式，再输入向量知识库中，由实验结果可知我们的模型和人类的阅读顺序的性能表现最佳。这是因为人类的阅读顺序可以极大的减少转换时的文字阅读顺序的丢失。从而可以增加向量数据库的语义连贯性和对应问题的质量。例如多栏论文如果不应用人类阅读顺序，在转为 markdown 格式时会出现大量的阅读顺序偏差问题。所以人类的阅读顺序和 GRPO 获得的阅读顺序都可以增强 RAG 系统的 QA 质量，并且可以看出人类的阅读顺序可以显著的提高大模型 F1 的值，这是由于默认的阅读顺序和规则式的阅读时顺序随着页数的增加，更加容易出现下一页时文字发生截断和乱序的情况，所以我们使用人类的阅读顺序可以有效的优化这一点。

表 4.5　文档问答任务结果，其中加粗的是最好结果，下划线是次好结果

Model	**PREORDER**	**LongDocURL**		**M6Doc**	**DocVQA**
		F1 ↑	ANLS ↑	ANLS ↑	ANLS ↑
LLama3-8b	OCR	40.06	41.12	72.77	67.21
	XYCUT	35.92	39.57	74.32	63.87
	Z-order	39.06	43.07	75.10	**68.44**
	Human	38.07	45.72	78.89	67.60
	Ours	**43.02**	**46.03**	**79.12**	67.56
GPT-4o	OCR	67.43	68.48	84.20	72.75
	XYCUT	64.91	66.97	80.21	70.34
	Z-order	63.66	71.49	88.77	74.61
	Human	**69.21**	73.68	89.03	74.45
	Ours	66.35	**74.13**	**89.69**	**74.88**
Qwen2-7B	OCR	60.18	64.66	78.83	65.42
	XYCUT	61.16	64.79	78.18	66.02
	Z-order	64.55	66.02	80.59	**67.59**
	Human	63.51	68.41	**80.90**	66.77
	Ours	**64.88**	**69.35**	80.87	67.47
GLM4v-9B	OCR	57.04	61.91	75.40	61.89
	XYCUT	58.74	60.57	75.76	62.55
	Z-order	55.43	62.44	**77.18**	65.41
	Human	59.41	64.12	76.82	65.22
	Ours	**60.36**	**64.30**	77.01	**65.47**

其次效果次优的是 Z 字顺序，Z 字顺序是大模型大量预训练时使用的通用阅读顺序，所以使用 Z 字顺序可以很好的激发出大模型的理解能力，所以在 LongDocURL 中表现也是较好的。

同时在 M6Doc 数据集中，效果类似，但是这里的增强不如在 LongDocURL 的增强，这是因为这个数据集的长度较为平均，并且都是较短的文档数据，所以在这个数据集的表现没有 LongDocURL 那么惊艳。并且在这个数据集中 Z 字顺序获得了最优的效果，这是因为在中短分布较为规律的文档中，Z 字顺序足够获得较好的表现结果。同时人类的阅读顺序同样可以让 LLMs 获得更加优质的效果。增强他们的 QA 理解能力。

最后在 DocVQA 数据集中，我们的方法效果和 Z 字顺序相差不大，这个是因为在单页的文档中，Z 字顺序也可以很好的提取出里面的信息内容，并且人类的阅读顺序对他的理解同样也是起到促进作用。

4.5.5 命名实体识别任务

我们的阅读顺序抽取大模型具备一定的迁移学习能力，可以很好的通用抽取文档的阅读顺序。我们使用的 M6Doc 数据集中，存在有大量的具有粗粒度和细粒度标签的标注数据，我们同样适用相同的大模型来分别进行粗粒度和细粒度的 SER 实验。分别从实验效果中进行分析，融合了人类阅读顺序的 SER 的效果会优于其他的阅读顺序。这是因为人类的阅读顺序可以增强 LLMs 的理解能力，增强 LLMs 查询实体的能力。

其次是 Z 字顺序，同理 Z 字顺序对 SER 的增强也是非常的大，大量的预训练使用 Z 字顺序为 LLMs 带来了海量的先验知识。并且 Qwen 模型取得了更大的增强，这说明 Qwen 模型具备更强的理解增强能力。

4.5.6 可视化分析

为了验证所提出模型在实际文档阅读顺序抽取中的效果，我们对 LongDocURL、M6Doc 以及 DocVQA 三个具有代表性的文档数据集进行了可视化展示，如图4.7所示。通过比较模型生成的阅读顺序与人类标注的真实阅读顺序，可以观察到我们的方法在绝大多数场景中表现出更高的一致性与稳定性。具体而言，模型输出的阅读路径整体上呈现出清晰的“从左至右、从上到下”的自然阅读流动，避免了传统方法

表 4.6　大模型直接进行迁移学习结果展示，粗体表示最优，下划线表示次优

Model	预排序方法	粗粒度 M6Doc			细粒度 M6Doc		
		P ↑	R↑	F1↑	P↑	R↑	F1↑
LLama3-8B	OCR	47.52	24.59	32.41	16.82	10.45	12.89
	XYCUT	45.36	26.67	33.59	9.48	12.61	10.82
	Z-order	**50.09**	28.44	36.28	**20.09**	11.84	14.90
	Ours	48.31	**29.06**	**36.29**	19.30	**13.33**	**15.77**
Qwen2-7B	OCR	68.09	36.92	47.88	43.51	26.71	33.51
	XYCUT	67.07	37.75	46.24	**50.22**	25.11	33.48
	Z-order	**69.49**	38.07	49.19	49.26	28.55	36.15
	Ours	66.86	**39.51**	**49.67**	49.43	**29.28**	**36.83**
GLM4-9B	OCR	70.14	39.92	50.88	47.35	29.71	36.51
	XYCUT	69.09	38.42	49.38	46.13	28.21	35.01
	Z-order	71.23	41.42	52.38	48.60	31.21	38.01
	Ours	**72.36**	**42.92**	**53.88**	**49.88**	**32.71**	**39.51**
GPT-4o	OCR	75.44	58.72	66.04	48.84	38.33	42.95
	XYCUT	71.18	60.78	65.57	49.03	40.57	44.40
	Z-order	75.61	**62.63**	68.51	46.05	**43.24**	44.60
	Ours	**77.50**	61.68	**68.69**	**52.06**	42.36	**46.71**

中常见的阅读顺序跳跃、不连贯等问题。这一性能的提升得益于预训练大模型中蕴含的大量通用先验知识，使其在面对复杂文档结构时，能够更好地模拟人类的阅读行为。并且我们的方法可以摒弃掉人类的一些阅读的习惯，例如突然的眼跳行为等，这样的阅读顺序相对于人类的阅读顺序而言会更加的稳定。

特别地，在如图4.7所示的双栏多页文档排版任务中，本模型依然能够较为准确地还原人类的阅读路径。模型不仅能够识别文档中的多栏结构，而且能够按照报刊常见的栏间阅读顺序，依次进行信息抽取，展现出良好的结构适应能力。这进一步验证了我们通过引入人类阅读顺序标注数据，对大模型进行风格诱导的有效性。实验结果表明，经由人类阅读行为标注训练的大模型能够学会更自然、符合认知规律的阅读策略，并在下游文档理解任务中显著提升整体表现。

4.5.7　消融实验

为了进一步分析不同类型奖励信号对模型性能的影响，我们设计了针对强化学习阶段所引入的两种奖励机制人类对齐奖励与格式化奖励的消融实验。具体地，我

图 4.7　通用阅读顺序抽取大模型结果展示

表 4.7　消融实验结果, 粗体表示最优，下划线表示次优

模型	**奖励函数类型**	**LongDocURL**		**M6Doc**	**DocVQA**
		F1 ↑	ANLS↑	ANLS↑	ANLS ↑
LLama3-8B	人类对齐奖励 + 格式化奖励	**43.02**	**46.03**	**79.12**	**67.56**
	人类对齐奖励	41.25	44.64	77.05	65.03
	格式化奖励	38.85	42.37	74.76	63.88
Qwen2-7B	人类对齐奖励 + 格式化奖励	**64.88**	**69.35**	**80.87**	**67.47**
	人类对齐奖励	62.10	67.91	77.03	66.08
	格式化奖励	60.25	65.93	76.38	63.54
GLM4-9B	人类对齐奖励 + 格式化奖励	**63.36**	**64.30**	**77.01**	**65.47**
	人类对齐奖励	60.43	62.97	76.39	64.67
	格式化奖励	59.61	59.57	73.77	60.52
GPT4o	人类对齐奖励 + 格式化奖励	**66.35**	**74.13**	**89.69**	**74.88**
	人类对齐奖励	64.91	72.82	87.89	73.18
	格式化奖励	62.12	68.73	84.98	71.54

们在 LongDocURL、M6Doc 以及 DocVQA 三个数据集上分别移除上述两类奖励函数，对比在不同奖励组合下模型在下游任务中的表现变化，实验结果如表4.7所示。

实验结果表明，当同时使用人类对齐奖励与格式化奖励时，模型在各项评估指标上取得了最优性能。这表明模型在优化过程中既能够学习符合人类阅读习惯的顺序，又能够输出结构简洁、信息集中的答案，从而有效缓解大模型可能出现的幻觉问题[86-87]，提升其对任务核心内容的感知与抽取能力。

当移除人类对齐奖励，仅保留格式化奖励时，模型的阅读顺序逐渐退化为简单的 Z 字型排序，缺乏对复杂文档结构的适应性，导致下游任务表现大幅下降。这说明人类对齐信号在引导模型模拟真实阅读行为方面具有关键作用。相反，当移除格式化奖励，仅保留人类对齐奖励时，虽然阅读顺序整体合理，但生成内容中存在大量无关信息，导致信息提取精度降低。尤其在长文档中，无效内容的混入进一步加剧了任务的困难度。

综上所述，两种奖励机制在强化学习阶段均发挥着互补作用：人类对齐奖励强化了阅读路径的合理性与任务对齐性，而格式化奖励则提升了模型输出的内容质量与可用性。二者协同作用，为大模型在多模态文档阅读与理解任务中提供了更具实用性的行为策略。

4.6 本章总结

本章主要针对当前多页文档阅读顺序数据集稀缺的问题，开展了系统性的数据标注与建构工作。我们以人类阅读过程中的眼动轨迹为依据，设计并实施了多种长度与排版结构的文档阅读顺序标注任务，涵盖了从单页到长篇幅的多页文档，最终构建了一个覆盖广泛、结构丰富的多页文档眼动阅读顺序数据集。

在数据集构建完成后，我们进一步将其应用于多模态大模型的训练与微调过程中。通过引入人类标注的阅读顺序作为监督信号，引导模型对其阅读策略进行优化，从而实现模型行为与人类阅读习惯之间的对齐。实验结果表明，经过该数据集微调后的多模态大模型不仅能够生成更符合人类直觉的阅读路径，并且可以摒弃掉人类阅读时会发生的而且在多个下游任务（如文档问答、信息抽取等）中均展现出更优异的性能表现。

综上所述，本章的工作有效弥补了多页文档阅读顺序数据资源的空白，为后续多模态文档理解任务提供了重要的数据支撑和模型训练基础。同时，也验证了通过引入人类眼动阅读行为实现模型对齐的可行性与有效性。

第五章　总结和展望

5.1　总结

在信息化时代视觉富文档（VRDs）应用日益深入的背景下，本文探讨了提升人工智能模型理解此类文档的关键路径：赋予 AI 自主探索与决策阅读顺序的能力。视觉富文档融合了文本、版式和视觉元素，其布局结构本身蕴含重要语义关联。然而，现有模型常因难以主动解析并适应 VRDs 复杂的布局多样性而理解受限。人类读者理解 VRDs 的优势，核心在于能动态规划阅读路径并分配视觉注意力。

因此，本文的核心在于研究如何使人工智能模型能够自主地探索视觉富文档的布局结构，模拟人类认知机制中的阅读顺序决策过程。我们聚焦于开发模型内在的布局感知与路径规划能力，使其能够像人类一样，依据文档的视觉和结构线索，主动地、策略性地确定信息获取的先后顺序和焦点，从而实现对 VRDs 更深层次、更高效的理解。本文的主要工作和贡献即围绕这一人工智能自主探索阅读顺序与注意力机制在视觉富文档理解中的应用展开。本文的贡献有以下几个方面：

贡献一：提出了强化学习根据下游任务自主探索文档阅读顺序的方法为了解决现有的规则式阅读顺序僵化，人类标注阅读顺序成本高昂，迁移性差的问题，我们提出了 HiDReader，一种利用强化学习探索-奖励的过程对阅读行为进行建模，可以有效的根据下游任务的反馈来动态调整文档阅读顺序的方法. 另外本文通过迁移学习的方式证明了文档的阅读模式在一定程度上具备特点即全局有序，局部不必完全有序。实验结果表明，这样的阅读模式可以很好的提高基础模型的表现性能。

贡献二：提出了多页文档阅读顺序数据集，并且根据该数据集标注得到了阅读顺序抽取的大模型

为了解决多页文档理解任务中跨页信息截断缺失的问题，我们对大量长文档进行了阅读顺序的标注，并且根据该阅读顺序对 VLM 进行人类偏好对齐训练。从而可以有效的提取出类似于人类阅读顺序的文档阅读顺序，并且在下游实验中验证得到人类的阅读顺序对长文档的理解任务中有着重要的促进作用。

5.2 展望

本文针对视觉富文档的阅读顺序抽取任务进行了研究并提出了相应的解决方案，但是依然存在很多的不足之处和可以改进之处。

(1) 自主探索阅读顺序的奖励函数优化

本文使用下游任务的回报作为奖励函数的核心，为了稳定的让模型可以很好的规避无效的探索空间，本文还使用了相似度奖励函数作为稳定模型表现的奖励函数，但是这样会限制模型进行有效探索的频率，所以我们还需要一个可以诱导模型进行探索行为的奖励函数，以及为了规避下游任务的奖励变化过于频繁和不稳定，我们还可以使用 GRPO 的思想对模型的更新进行拓展，让模型可以在一个分组内进行阅读顺序的更新和调整。此外，我们提出的方法需要训练大量的时间，在训练过程中需要对超参进行调整从而让模型可以获得较好的表现能力，后续也需要对训练过程进行缓存和优化。

(2) 提升阅读顺序抽取大模型的能力

当前对于长文档阅读顺序的研究仍处于相对初级阶段，尤其在面对超长篇幅的文档，如整本书籍、法规文档或技术手册等场景时，现有模型的阅读顺序提取能力仍显不足。一方面，随着文档页数的大幅增加，依赖人工进行跨页阅读顺序标注的成本急剧上升，严重制约了大规模高质量数据集的构建。另一方面，长文档中的阅读行为往往跨越多个页面，呈现出非线性、跳跃式的逻辑关系，这对模型的记忆能力、全局感知能力以及上下文建模能力提出了更高的要求。

因此，提升大模型在超长文档中的阅读顺序建模能力成为亟需解决的问题。下一阶段的研究方向将聚焦于两个方面：其一，进一步扩充覆盖多页、跨结构的阅读顺序标注数据，引入更多具有人类眼动特征的监督信息，从而丰富模型的感知能力和对复杂结构的理解能力；其二，探索利用少量高质量阅读顺序数据，通过指令微调、迁移学习或强化学习等方式，激发多模态大模型对阅读路径的启发式涌现能力，增强其泛化能力和任务适应性。

插图索引

表格索引

参考文献

[1] 崔磊, 徐毅恒, 吕腾超, 等. 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] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.

[3] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818.

[4] XU Y, LI M, CUI L, et al. Layoutlm: Pre-training of text and layout for document image understanding[C]//KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2020.

[5] XU Y, XU Y, LV T, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding[Z]. 2020.

[6] HUANG Y, LV T, CUI L, et al. Layoutlmv3: Pre-training for document ai with unified text and image masking[C]//2022.

[7] LI Y, QIAN Y, YU Y, et al. Structext: Structured text understanding with multi-modal transformers[C]//2021.

[8] ZHAI M, LI Y, QIN X, et al. Fast-structext: An efficient hourglass transformer with modality-guided dynamic token merge for document understanding[Z]. 2023.

[9] LYU P, LI Y, ZHOU H, et al. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond[Z]. 2024.

[10] PEI X, GUO S, HU Y, et al. Graph pointer network assisted deep reinforcement learning for virtualized network embedding[J]. IEEE Transactions on Green Communications and Networking, PP.

[11] YIN Y, MENG F, SU J, et al. Enhancing pointer network for sentence ordering with pairwise ordering predictions[C]//2020.

[12] WANG Z, XU Y, CUI L, et al. Layoutreader: Pre-training of text and layout for reading order detection[Z]. 2021.

[13] WANG H, WANG Q, LI Y, et al. Doctrack: A visually-rich document dataset really aligned with human eye movement for machine reading[Z]. 2023.

[14] 丁世飞, 杜威, 张健, 等. 多智能体深度强化学习研究进展[J]. 计算机学报, 2024, 47(7): 1547-1567.

[15] SHAO Z, WANG P, ZHU Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models[Z]. 2024.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[17] ZHANG C, GUO Y, TU Y, et al. Reading order matters: Information extraction from visually-rich documents by token path prediction[Z]. 2023.

[18] 仵博, 吴敏. 部分可观察马尔可夫决策过程研究进展[J]. 计算机工程与设计, 2007, 28(9): 2116-2119.

[19] 章倩, 王梓祺. 基于自定义模板的 OCR 技术及应用[J]. 指挥信息系统与技术, 2023, 14(5): 94-98.

[20] CHEN X, JIN L, ZHU Y, et al. Text recognition in the wild: A survey[J]. ACM Computing Surveys, 2022, 54(2).

[21] DAVIS R A, LII K S, POLITIS D N. Remarks on some nonparametric estimates of a density function[M]//2011.

[22] HOTI F. On estimation of a probability density function and mode[Z]. 2003.

[23] MIRMEHDI M, XIE X, SURI J. Handbook of texture analysis[M]. Handbook of Texture Analysis, 2008.

[24] KIM Y. Convolutional neural networks for sentence classification[J]. Eprint Arxiv, 2014.

[25] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[M]//International Conference on Learning Representations. 2015.

[26] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[A]. 2014.

[27] GRAVES A. Long short-term memory[J]. Springer Berlin Heidelberg, 2012.

[28] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C]//Vol. 38. 2013.

[29] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//abs/1810.04805. 2019.

[30] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[Z]. 2018.

[31] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2017, 39(6): 1137-1149.

[32] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[33] SUN Y, WANG S, LI Y, et al. Ernie: Enhanced representation through knowledge integration[Z]. 2019.

[34] SUN Y, WANG S, LI Y, et al. Ernie 2.0: A continual pre-training framework for language understanding[C]//2020: 8968-8975.

[35] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [Z]. 2020.

[36] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[Z]. 2019.

[37] CHEN M, TWOREK J, JUN H, et al. Evaluating large language models trained on code[A]. 2021.

[38] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[39] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[A]. 2023.

[40] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[A]. 2023.

[41] DRIESS D, XIA F, SAJJADI M S, et al. Palm-e: An embodied multimodal language model[Z]. 2023.

[42] DUBEY A, JAUHRI A, PANDEY A, et al. The llama 3 herd of models[A]. 2024: arXiv-2407.

[43] TEAM G, GEORGIEV P, LEI V I, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[A]. 2024.

[44] LIU A, FENG B, XUE B, et al. Deepseek-v3 technical report[A]. 2024.

[45] GUO D, YANG D, ZHANG H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[A]. 2025.

[46] PANG B, LI Y, LI J, et al. Tdaf: Top-down attention framework for vision tasks[C]// Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 35. 2384-2392.

[47] RITTER F E, TEHRANCHI F, OURY J D. Act-r: A cognitive architecture for modeling cognition[J]. Wiley Interdisciplinary Reviews: Cognitive Science, 2019, 10(3): e1488.

[48] WOLFE J M, CAVE K R, FRANZEL S L. Guided search: an alternative to the feature integration model for visual search.[J]. Journal of Experimental Psychology: Human perception and performance, 1989, 15(3): 419.

[49] KATSUKI F, CONSTANTINIDIS C. Bottom-up and top-down attention: different processes and overlapping neural systems[J]. The Neuroscientist, 2014, 20(5): 509-521.

[50] GU Z, MENG C, WANG K, et al. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding[C]//2022.

[51] JAUME G, EKENEL H K, THIRAN J P. Funsd: A dataset for form understanding in noisy scanned documents[C]//2019 International Conference on Document Analysis and Recognition Workshops (ICDARW): Vol. 2. IEEE, 2019: 1-6.

[52] MATHEW M, BAGAL V, TITO R, et al. Infographicvqa[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 1697-1706.

[53] LI Q, LI Z, CAI X, et al. Enhancing visually-rich document understanding via layout structure modeling[C]//Proceedings of the 31st ACM international conference on multimedia. 2023: 4513-4523.

[54] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in neural information processing systems, 2020, 33: 9459-9474.

[55] DENG C, YUAN J, BU P, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating[A]. 2024.

[56] CHENG H, ZHANG P, WU S, et al. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 15138-15147.

[57] SUTTON R S, BARTO A G, et al. Reinforcement learning: An introduction: Vol. 1 [M]. MIT press Cambridge, 1998.

[58] 夏乐天, 朱元生. 马尔可夫链预测方法的统计试验研究[C]//2007 重大水利水电科技前沿院士论坛暨首届中国水利博士论坛论文集. 2007.

[59] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//International conference on machine learning. PmLR, 2016: 1928-1937.

[60] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation[A]. 2015.

[61] WU Y, MANSIMOV E, GROSSE R B, et al. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation[J]. Advances in neural information processing systems, 2017, 30.

[62] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[A]. 2017.

[63] HEESS N, TB D, SRIRAM S, et al. Emergence of locomotion behaviours in rich environments[A]. 2017.

[64] 赵维森. 视觉文化时代人类阅读行为之嬗变[J]. 学术论坛, 2003(3): 127-131.

[65] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[A]. 2020.

[66] BAI J, BAI S, CHU Y, et al. Qwen technical report[A]. 2023.

[67] BAI J, BAI S, YANG S, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond[Z]. 2023.

[68] DU Z, QIAN Y, LIU X, et al. Glm: General language model pretraining with autoregressive blank infilling[A]. 2021.

[69] TU Y, GUO Y, CHEN H, et al. Layoutmask: Enhance text-layout interaction in multimodal pre-training for document understanding[Z]. 2023.

[70] DEHLING H, VOGEL D, WENDLER M, et al. Testing for changes in kendall's tau [J]. Econometric Theory, 2017, 33(6): 1352-1386.

[71] SEDGWICK P. Spearman's rank correlation coefficient[J]. Bmj, 2014, 349.

[72] JIN Y, DING Y, PAN X, et al. Pointerformer: Deep reinforced multi-pointer transformer for the traveling salesman problem[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 37. 2023: 8132-8140.

[73] ZHANG C, TU Y, ZHAO Y, et al. Modeling layout reading order as ordering relations for visually-rich document understanding[Z]. 2024.

[74] CECI M, BERARDI M, PORCELLI G A, et al. A data mining approach to reading order detection[C]//International Conference on Document Analysis Recognition. 2007.

[75] VILLANOVA-APARISI D, TARRIDE S, MARTÍNEZ-HINAREJOS C D, et al. Reading order independent metrics for information extraction in handwritten documents[C]//International Conference on Document Analysis and Recognition. Springer, 2024: 191-215.

[76] SAHOO P, SINGH A K, SAHA S, et al. A systematic survey of prompt engineering in large language models: Techniques and applications[A]. 2024.

[77] WHITE J, FU Q, HAYS S, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt[A]. 2023.

[78] XIAO Z, YAN S, HONG J, et al. Dynaprompt: Dynamic test-time prompt tuning[A]. 2025.

[79] PENG Z, WANG W, DONG L, et al. Kosmos-2: Grounding multimodal large language models to the world[A]. 2023.

[80] CHEN K, ZHANG Z, ZENG W, et al. Shikra: Unleashing multimodal llm's referential dialogue magic: abs/2306.15195[A]. 2023.

[81] YOU H, ZHANG H, GAN Z, et al. Ferret: Refer and ground anything anywhere at any granularity[Z]. 2023.

[82] LI Z, XU Q, ZHANG D, et al. Groundinggpt:language enhanced multi-modal grounding model[Z]. 2024.

[83] XU J, LIU X, WU Y, et al. Imagereward: Learning and evaluating human preferences for text-to-image generation[Z]. 2023.

[84] CHRISTIANO P, LEIKE J, BROWN T B, et al. Deep reinforcement learning from human preferences[Z]. 2017.

[85] RAPOSO D, RITTER S, RICHARDS B, et al. Mixture-of-depths: Dynamically allocating compute in transformer-based language models[Z]. 2024.

[86] RAWTE V, SHETH A, DAS A. A survey of hallucination in large foundation models [Z]. 2023.

[87] DHULIAWALA S, KOMEILI M, XU J, et al. Chain-of-verification reduces hallucination in large language models[Z]. 2023.

作者在攻读硕士学位期间发表的论文与研究成果

一. 论文

[1]2025.HiDReader: Human-Inspired Document Reading Agent via Reinforcement Learning. The 19th International Conference on Document Analysis and Recognition (ICDAR 2025). (第一作者，CCF-C，*oral*)

[2]2023. DocTrack: A Visually-Rich Document Dataset Really Aligned with Human Eye Movement for Machine Reading. In Findings of EMNLP 2023.(第四作者)

致　谢

首先感谢王昊老师三年来的指导，是王老师带领我们入门科研，进行理论教导，他严谨的科研态度，敏锐的科研直觉，强大的编程能力都让我终身受益匪浅，王老师严谨治学的思想对我未来的发展有着巨大的启发和提升。其次我要感谢朱频频老师的教导，朱老师教会了我如何利用在学校学习到的知识转化为工作需要的工程能力。

我要感谢上海大学给我提供了研究生的平台，让我有机会见到如此多学术能力超强，思想新颖治学一丝不苟的老师。我要感谢父母、同学等所有帮助过我的人，是你们在研究生三年与我交流了未来的发展方向，共同进行学术上能力的提升。最后我要感谢党和国家，为我们提供了安稳学习的环境和极好的工作环境社会氛围。

诚挚感谢各位审稿专家的辛勤付出。正是由于您们对论文质量的严格把关，才促使相关研究成果不断提升，也为我在撰写过程中提供了高质量的参考依据。

最后我要感谢华南理工大学的金连文老师对于我们实验数据集的支持，您严谨开放的科研风度，是我需要学习的风骨。