中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：21721586

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	跨模态提示增强的视觉富文档理解方法研究

作　　者　陈夏华
学科专业　计算机应用技术
导　　师　王昊
完成日期　二〇二四年五月

姓　　名：陈夏华　　学号：21721586

论文题目：跨模态提示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:　　年　　月　　日

姓　　名：陈夏华　　学号：21721586

论文题目：跨模态提示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:

日期:　　年　　月　　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:　　导师签名:

日期:　　年　　月　　日　　日期:　　年　　月　　日

上海大学工学硕士学位论文

跨模态提示增强的视觉富文档理解方法研究

作　　者: 陈夏华
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
2024 年 5 月

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

Research on Cross-Modal Prompt-Enhanced Visually-Rich Document Understanding Methods

Candidate: Xiahua Chen
Supervisor: Hao Wang
Major: Technology of Computer Application

School of Computer Engineering and Science
Shanghai University
May, 2024

摘　要

随着数字化时代的快速发展，视觉富文档作为一种新型的信息传播形式，正在受到越来越多的关注和应用。由于众多领域的文档版式多样，形式不一，采用人工的方式处理这些文档十分低效且昂贵。因此，文档智能模型的研究和自动化抽取技术的开发具有重要的应用价值。相比于传统的纯文本文档，视觉富文档具有复杂的版式布局和更多的视觉元素（包括颜色、字体大小、样式等），这些布局和视觉信息对于文档的认知和理解起着至关重要的作用。通常情况下，视觉富文档中的文本、布局和视觉蕴含着一致的语义信息，布局之间的关系可能会提示文本内容之间的关系，同样，视觉元素也会提示布局之间的关系，要充分利用这些关系需要对不同模态的信息进行融合和对齐。本文围绕视觉、排版布局、文本内容的跨模态对齐，所作的工作如下：

（1）**基于视觉非对称学习的文档实体语义表示方法**：现有的多模态视觉富文档理解模型大多是通过文本-图像，单词-单词块等粗粒度对齐的方式融入视觉特征，忽略了与文档排版布局相关的细粒度视觉特征，如背景颜色、字体、位置等。此外，之前的方法关注在文本和布局的建模，视觉通道的融合方式单一，导致多个模态之间的信息流无法均衡。这使得，第一，融入模型的视觉特征受限；第二，多模态模型学习的特征是有偏的。为了解决视觉特征受限的问题，本文提出一种基于颜色块的视觉提示方法，将不同类别的实体用不同的颜色进行填充，使模型更加关注实体所对应的图像块内容和边界，从而捕捉细粒度的视觉特征。为了解决有偏的多模态学习，本文提出一种基于视觉-排版先验的多模态一致性学习框架 VANCL。在注意力机制融合的基础上，额外增加了基于颜色先验的排版信息，通过一致性学习实现无偏多模态表示。实验表明，VANCL 在三个数据集和多个主干网络上都显著优于之前的多模态模型，达到目前最优的性能（SOTA）。

（2）**基于布局感知提示的大模型文档实体关系理解方法**：目前基于大语言模型的视觉富文档理解研究还不完善。虽然在超大规模文本语料上预训练的大语言模型，如 ChatGPT，已经在大量的纯文本任务上取得了颠覆性的成果，但是针对于需要考虑二维位置的视觉富文档理解任务，大语言模型的潜力还未被挖掘。

因此，本文提出一种多角度空间位置理解的评估方案，能够综合的评估大语言模型的文档理解能力和生成能力，该方案包括基本空间感知、页面位置预测、文档信息抽取和文档布局生成这四个方面，本文构建了与之对应的四个二维位置理解的评估数据子集，用于高效的评测不同的大语言模型。此外，本文提出一种基于文本-排版跨模态提示增强的文档实体关系理解方法，将含有版式布局的文本信息作为大语言模型的上下文范例，并引入键值匹配模式的布局提示。实验表明，提出的方法显著提升了文档智能模型的理解能力。

本文的研究成果对于文档智能模型的发展具有重要的意义。特别地，本文提出的大语言模型空间感知能力的评估方案和数据子集为大语言模型的版式感知研究提供了有力的研究基础。

关键词：视觉富文档；跨模态对齐；视觉提示；大语言模型评估；布局提示

ABSTRACT

With the rapid development of the digital era, Visually-Rich Documents(VRDs) have emerged as a novel form of information dissemination, garnering increasing attention and application. Due to the diversity of document formats across various fields, manual processing of these documents is both inefficient and costly. Hence, the study of Document AI Models and the development of automated extraction technologies hold significant practical value. Compared with traditional plain text documents, VRDs have complex layouts and more visual elements, including color, font size, style, etc., which play a crucial role in the cognition and understanding of the document. Typically, the text, layout, and visual elements in VRDs carry consistent semantic information. The relationship between layouts may indicate the relationship between text content. Similarly, visual elements may also indicate the relationship between layouts. To fully leverage these relationships, it is essential to integrate and align information across different modalities. This paper focuses on the cross-modal alignment of visuals, layouts, and textual contents. The work done is as follows:

(1) **Document entity semantic representation based on visual asymmetric learning**: Most of the existing multi-modal VRDs understanding models integrate visual features through coarse-grained alignment such as text-image, word-patch, etc., ignoring fine-grained visual features related to document layout, such as background color, font, position. In addition, previous methods focus on the modeling of text and layout, and the fusion method of visual channels is single, resulting in an unbalanced information flow among multiple modalities. This makes, first, the visual features integrated into the model limited; second, the features learned by the multi-modal model are biased. In order to solve the problem of limited visual features, this paper proposes a visual prompt method based on color patch, which fills different categories of entities with different colors, so that the model pays more attention to the content and boundaries of the image patch corresponding to the entities, thereby capturing fine-grained

visual features. In order to solve the biased multi-modal learning, this paper proposes a multi-modal consistency learning framework based on visual-typesetting prior VANCL. On the basis of the fusion of attention mechanisms, additional layout information based on color priors is added to achieve unbiased multi-modal representation through consistency learning. Experiments show that VANCL is significantly better than previous multi-modal models on three datasets and multiple backbone networks, reaching the current optimal performance (SOTA).

(2) **Document entity relationships understanding for large language models based on layout-aware prompts**: At present, research on VRDs understanding based on Large Language Models(LLMs) is not yet complete. Although LLMs pre-trained on very large-scale text corpus, such as ChatGPT, have achieved disruptive results on a large number of pure text tasks, LLMs potential has not yet been excavated for VRDs understanding tasks considering two-dimensional positions. Therefore, this paper proposes an evaluation scheme for multi-angle spatial position understanding, which can comprehensively evaluate the document understanding and generation capabilities of LLMs. The solution includes four aspects: basic spatial perception, page location prediction, document information extraction and document layout generation. This paper constructes four corresponding two-dimensional position understanding evaluation data subsets for efficient evaluation of different LLMs. In addition, this paper also proposes a method for understanding document entity relationships based on text-layout cross-modal prompt enhancement, using input containing layouts as context demonstrations and the layout prompt of the key-value matching pattern is introduced for LLMs. Experiments show that the proposed method significantly improves the understanding ability of Document AI Model.

The research results of this paper have certain significance for the development of Document AI Models. In particular, the evaluation scheme and data subset proposed in this paper provide a strong research foundation for layout awareness of LLMs.

Keywords: Visually-rich documents; cross-modal alignment; visual prompts; large language model evaluation; layout prompts

目　录

第一章　绪论

1.1　研究背景及意义

随着数字化时代的快速发展，信息量不断爆炸，人们对信息的获取和处理需求日益增加。在这个背景下，传统的纯文本信息已经不能完全满足人们的需求，而视觉富文档作为一种新型的信息传播形式，正在受到越来越多的关注和应用。视觉富文档[1]是一种将文本、图片、图表、视频等多种形式的信息内容有机融合在一起的文档形式。相比于传统的纯文本文档，视觉富文档具有更多的视觉元素（包括颜色、字体大小、文本位置、样式等），这种视觉信息对于文档的认知和理解起着至关重要的作用。

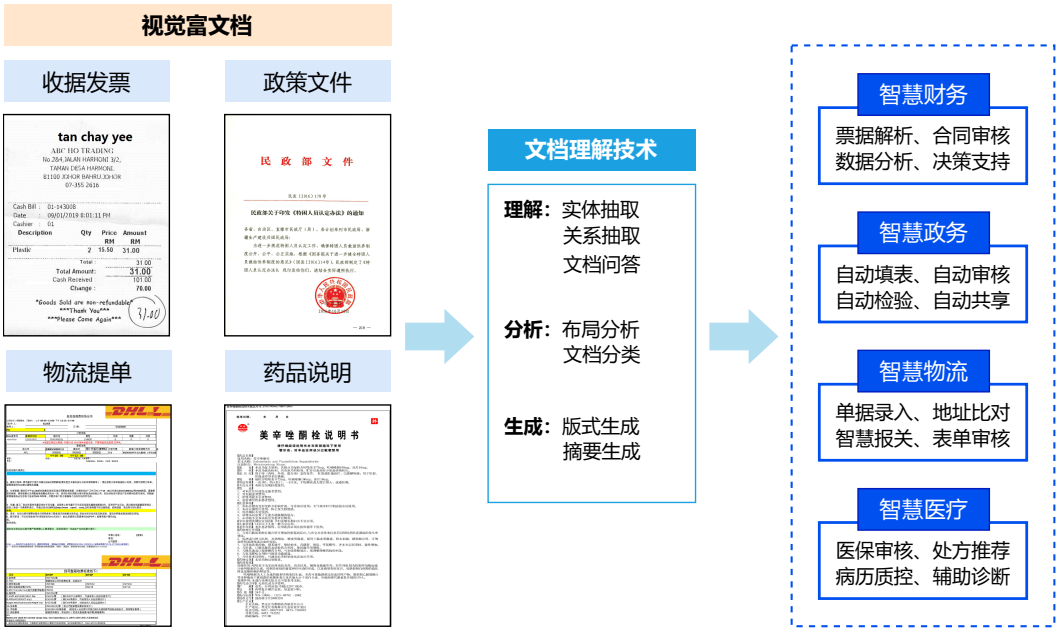


图 1.1　视觉富文档智能理解技术的应用场景

在大数据时代，各个领域人工处理海量的视觉富文档，虽然有着很高的准确度，但是费时费力，并且需要高昂的人工成本。因此，针对视觉富文档的智能理解技术应运而生。目前，文档智能理解技术广泛应用于金融、政务、法律、物流、医疗等行业，常见的应用场景包括财务报销单、招聘简历、企业财报、合同文书、政策文件、法律判决书、物流单据等多模态文档的关键信息抽取、文档解

析、文档比对等，如图 1.1 所示。然而，正因为众多领域的文档版式多样，形式不一（PDF，Word，PNG 等），采用人工的方式处理这些文档十分低效且昂贵。以物流单据为例，现有的公司大多采用规则和模板匹配辅以人工的方式处理货运表单，这种方式适用的表单布局单一，不具备扩展性，并且在自动化的流程中出现了严重的效率瓶颈。因此，视觉富文档智能理解模型（以下简称文档智能模型，Document AI Model）的研究和自动化抽取技术的开发不仅可以提高文档处理的效率和准确性，还可以节约大量的人力成本，这具有重要的应用价值。

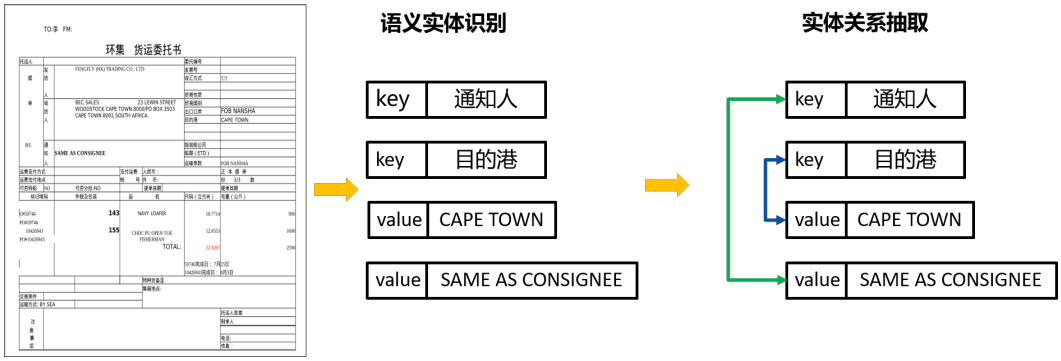


图 1.2　视觉富文档的关键信息抽取过程

然而，相对于纯文本文档而言，视觉富文档具有版式结构复杂、视觉特征多样的特点，并且由于文档扫描图像质量差、文档形式不统一等问题，导致视觉富文档的语义理解十分具有挑战。目前基于纯文本的理解主要是通过信息抽取这个核心任务来进行的，如图 1.2 所示，信息抽取[2]是一种从无结构和半结构化的海量数据中抽取结构化知识的技术。在纯文本中，信息抽取的主要任务是从过单词、句子和篇章的语义信息抽取文本中的关键信息，如命名实体、实体关系等。然而，由于视觉富文档的结构松散，信息密度小，仅仅通过纯文本进行信息抽取难以捕捉深度的语义信息，视觉富文档中丰富的视觉元素，版式特征能够极大地帮助机器理解文档的语义。因此，视觉富文档理解的研究要考虑更多的版式与视觉的特征，这一项研究还需进一步的探索。

1.2　研究问题

本文研究的对象是视觉富文档，研究的主体的文档智能模型，包含传统的多模态视觉富文档理解模型以及大语言模型。目前的多模态视觉富文档理解模型

需要针对特定的文档理解任务，构建特定的数据集和训练特定的模型，而大语言模型利用通用提示模板的方式将多个理解任务统一起来，进一步提升了文档模型的智能化。本文旨在提升传统多模态视觉富文档模型的理解能力，并探索大语言模型对于视觉富文档的理解能力，这需要围绕文本、布局、视觉三个模态进行深入研究。因此，本文主要关注以下两个研究问题：

（1）如何利用视觉-排版跨模态提示提升多模态视觉富文档模型的理解能力

视觉富文档中包含了众多的视觉元素，包括背景颜色、边界框线、文字风格等。对于关键信息抽取任务来说，文档中的关键信息通常以键（key）值（value）对的形式出现，这里的键和值往往会有视觉层面的差异，比如键（key）常常以加粗、大号的字体出现，值（value）通常没有明显的视觉特征，因此键值之间的视觉差异能够很容易的把键-值区分。除此以外，文档标题的视觉信息与其他正文文本的信息也有较大差异，文本是否被加粗，是否有底色都能说明该文本的重要性。除视觉信息外，文档的排版布局在理解语义时也是不可或缺的，布局信息在文档中通常是以文本所在的边界框（bounding box）的形式存在，而边界框由左上、右下两个坐标点唯一确定。在判断键值关系时，键和值的相对位置关系能帮助机器快速理解键和值的语义关系。例如，在发票收据的数据中，“总金额”通常会出现在文档的右下部分作为结算，而总金额具体的数字，会紧接的出现在“总金额”右方或者下方，这种显著的位置提示在纯文本的序列模型中被忽视，导致模型性能受限。

先前的多模态视觉富文档理解模型（Large-scale pretrained Multi-modal Models，LMMs）[3-4]利用的视觉特征是从粗粒度对齐的角度入手的，如文本-图像（text-image）、单词-单词块（word-patch）等，而视觉的细粒度特征对文档语义的理解也尤为重要。目前的文档智能模型大多基于 OCR 引擎和 Transformer 架构，难以直接从粗粒度的视觉特征中学习细粒度的视觉特征。因此，如何有效利用粗细粒度的视觉特征是需要解决的问题。

此外，之前的方法[5]过于的关注在文本和布局建模，视觉通道的融合方式单一，导致多个模态之间的信息流无法均衡。视觉-排版之间的信息交互可以有效地提升多模态表示的稳定性，如何采用一种新的视觉-排版融合方式是需要探索的。

（2）如何利用文本-排版跨模态提示提升大语言模型的版式感知能力

传统的视觉富文档理解模型（LMMs）通过海量视觉富文档的预训练来学习多模态表示，有比较强的领域壁垒，虽然可以通过标注少量其他领域的数据微调来解决领域迁移问题，但还是需要繁琐的数据标注，以及繁琐的流程操作（pipeline）。从 ChatGPT 诞生以来，大语言模型（Large Language Models，LLMs）凭借简单的自然语言交互和强大的零样本推理能力迅速在各个自然语言处理（NLP）任务中取得最先进的成果。ICL-D3IE[6] 已经证明了 LLMs 具备一定的空间位置理解能力，文档智能模型（Document AI Model）是否从预训练语言模型迁移至大语言模型取决于其本身的二维空间位置理解能力。然而，大语言模型的训练语料是基于纯文本的，缺少与视觉富文档相关的二维文档，所以大语言模型对于二维文档的理解能力还需要进一步探究。

对于大语言模型而言，目前缺乏一套关于二维位置坐标理解、布局感知的评估方案，这套评估方案对于文档智能模型的版式感知研究具有重要的意义。此外，如何利用大语言模型强大的文本理解能力以及丰富的先验知识来帮助文档排版布局的理解，也需要进一步的探索。

1.3 研究内容

本文要解决文档智能模型中存在的多粒度视觉特征表示、多模态特征（文本、布局、视觉）不均衡、大模型二维空间位置理解能力评估和文本提示排版布局这四个问题，应当考虑：视觉-排版提示和文本-排版提示这两方面。如图 1.3 所示。

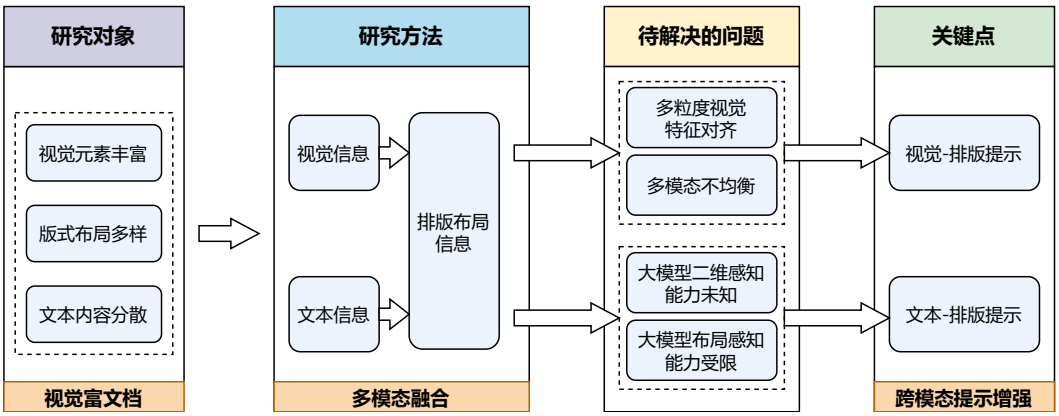


图 1.3　本文的研究内容

（1）针对如何利用视觉-排版提示来提升多模态视觉富文档模型的理解能力，需要解决两个问题：基于 Transformer 架构的文档智能模型难以学习细粒度视觉特征的问题、视觉通道的融合方式单一导致多个模态之间的信息流无法均衡的问题。为了解决第一个问题，本文提出一种基于颜色块（color patch）的视觉提示方法，将不同类别的实体用不同的颜色在图片上进行填充。通过训练一个新的视觉增强网络作为" 教师" 模型，该模型在训练过程中由于颜色块的视觉特征，更加关注实体所对应的图像块内容和边界，从而捕捉更细粒度的视觉信息。为了解决第二个问题，本文提出一种基于视觉-排版先验的多模态一致性学习框架，此框架在注意力机制融合多模态特征的基础上，额外增加了基于颜色块的视觉信息，通过一致性学习，将“教师”模型学习的细粒度特征迁移到“学生”模型，从而达到多模态均衡。大量的实验表明，本文所提出的方法显著优于之前的 LMMs，达到目前最优的性能（SOTA）。

（2）针对如何利用文本-排版提示来提升大语言模型的版式感知能力，需要解决两个问题：基于纯文本训练的大语言模型在视觉富文档上的二维空间位置理解能力未知的问题、二维空间的排版布局难以通过文本语义建模导致性能受限的问题。为了解决第一个问题，本文提出一套适用于大语言模型二维空间位置理解能力的评估方案，从四个方面评测大语言模型的位置理解能力，位置预测能力，二维键值匹配能力和版式布局生成能力。为了解决第二个问题，本文基于大语言模型的二维感知能力评估，提出一种布局感知提示的文档实体关系理解方法，将含有版式布局的文本信息作为大语言模型的上下文范例，并引入键值匹配模式的布局提示，有效的提升文档智能模型的版式感知能力。实验结果表明，本文提出的评估方案能客观的评测大语言模型的二维空间位置理解能力，由此引申的文本-排版跨模态提示方法能显著提升大语言模型的文档理解能力。

1.4 创新点

本文针对视觉富文档的语义实体理解任务，从多模态语义粒度对齐、文档智能模型的版式感知能力两方面进行研究，图 1.4 展示了本文的创新点。

（1）**提出视觉非对称学习的文档实体语义表示方法，实现更好的多模态信息的对齐融合**：现有的多模态视觉富文档理解模型融入视觉信息的方式比较单

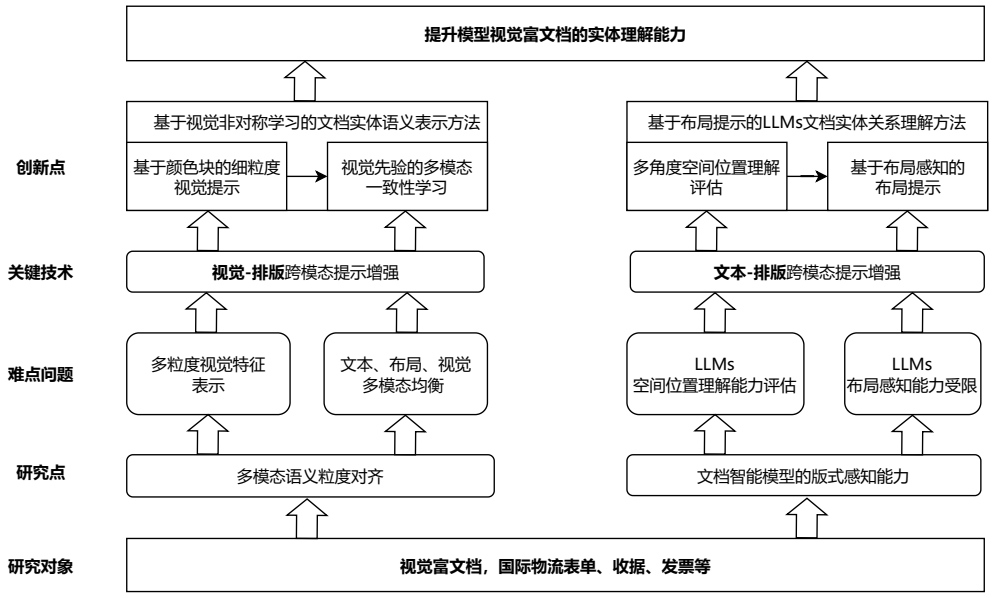


图 1.4　本文的创新点

一，例如 LayoutLM 模型[5]，仅仅在将视觉表示与多模态表示叠加，未考虑模态之间的交互。后续的改进模型也只考虑到粗粒度的视觉特征，如文本段-图像（text-image）和单词-单词块（word-patch）。此外，这些模型的视觉特征抽取都依赖预训练的视觉编码器，并没有再参与后续的训练。这使得，第一，预训练的视觉编码器无法捕捉与文档版式布局相关的细粒度视觉特征，比如文本边界框（bounding box）的位置，字体特征和颜色特征等。第二，预训练的视觉编码器不能很好的抽取二维文档的领域视觉特征。为了解决这些问题，本文提出一种基于颜色块的细粒度视觉提示方法和基于视觉-排版先验的多模态一致性学习框架。具体的说，文档中的每一个文本段都被填充与类别对应的颜色，形成颜色块（color patch），然后在训练阶段引入一个新的视觉编码器来学习带有颜色块的文档图像，这个过程可以看作是 OCR 的逆过程。在学习过程中，视觉编码器可以捕捉文本段检测和边界框回归的细粒度视觉特征，从而提升视觉编码器的能力。此外，本方法使用一致性学习将这个被增强的视觉编码器的能力迁移至文档智能模型，显著提升了文档智能模型的多模态理解能力。整个模型框架在不引入新的模型参数的前提下实现了非侵入式、可拔插、低人工。

（2）**提出布局感知提示的大模型文档实体关系理解方法，增强了大模型的版式感知能力**：大模型的版式感知要求模型具备空间位置的理解能力和排版布局的感知能力。目前基于大语言模型的视觉富文档理解研究还不完善。虽然在超大

规模文本语料上预训练的大语言模型，如 ChatGPT，已经在大量的纯文本任务上取得了颠覆性的成果，但是针对于需要考虑二维位置的视觉富文档理解任务，大语言模型的潜力还未被挖掘。因此，本文提出一种多角度空间位置理解的评估方案，能够综合的评估大语言模型的文档理解能力和生成能力。方案包括基本空间感知、页面位置预测、文档信息抽取和文档布局生成这四个方面，本文构建了与之对应的四个二维位置理解的评估数据子集，用于高效评测不同的大语言模型。此外，本文提出一种基于布局感知提示的文档实体关系理解方法，将含有版式布局的文本信息作为大语言模型的上下文范例，并引入键值匹配模式的布局提示。实验表明，该方法有效提升了文档智能模型的版式感知能力。

1.5 本文的组织结构

本文针对视觉富文档的跨模态提示增强开展深入的研究，首先介绍了视觉富文档的研究价值和研究内容，随后介绍了文档智能任务的定义和国内外研究的现状。针对现有的文档智能模型不能很好的利用视觉特征的问题，本文研究了文档智能模型的视觉增强方法。为了探究文档智能模型的版式感知能力，本文提出大模型二维空间位置理解的评估方法和布局提示策略，最后总结了本文的研究工作并进行展望。本文的组织结构如图 1.5 所示：

第一章为绪论，首先概述了视觉富文档理解的研究背景及其重要性，随后分析了当前视觉富文档理解领域中存在的两个核心问题，并据此提出了本研究旨在探讨的两个研究内容。最终，对本论文的创新点进行了总结。

第二章为视觉富文档信息抽取任务定义和相关方法概述。首先，本章对文档理解任务进行概述，介绍了文档智能模型的信息抽取技术框架。接着，分别从传统的多模态视觉富文档理解模型和大语言模型的角度梳理信息抽取方法的发展脉络，分析和比较了现有方法的特点和不足，为后续文本的工作做理论铺垫。

第三章介绍本文提出的 VANCL 方法。针对目前多模态模型难以捕捉细粒度视觉特征的问题，VANCL，通过颜色块增强视觉编码器的细粒度特征表示能力，并利用多模态一致性学习均衡多模态表示。在三个不同领域的视觉富文档数据集上进行了大量的实验验证方法的精确性和鲁棒性。

第四章探究文档智能模型的版式感知能力。本章主要围绕提出的四个二维

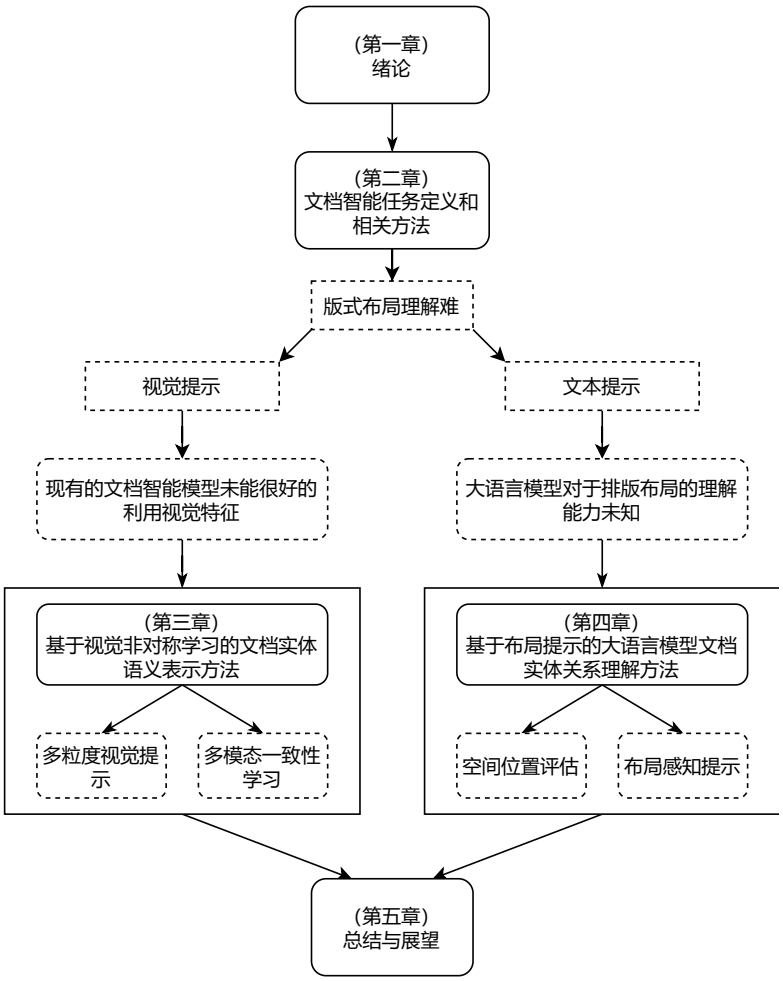


图 1.5　本文的组织结构

空间位置理解的评估方法，构建评估数据集，从多个角度评价大语言模型的排版空间理解和生成能力。通过比较参数量不同的大语言模型的评估结果和可视化图像，验证所提出方法的客观性。本章在四个评估方案的基础上提出布局感知提示的实体关系理解方法，显著提升了大语言模型的版式感知能力。

第五章对本文的研究工作进行总结，分析目前工作的不足之处和可拓展的方面，最后对视觉富文档理解未来的研究方向进行展望。

第二章　文档智能现有方法及模型

2.1　相关任务定义

视觉富文档理解（Visually-rich Document Understanding，VDU）技术[7]旨在从视觉富文档（Visually Rich Documents，VRDs）中归纳、整理、抽取有用的信息。日常生活中，视觉富文档的存在形式是多样的，如网页，Excel 表格，PDF，扫描图像等，这使得机器难以直接处理多种格式的 VRDs。现阶段大部分 VDU 模型一般使用两阶段方案解决这个问题：（1）从文档中读取文本信息（2）对文档文本进行全面的理解。第一个阶段，通常的做法是使用 HTML/XML 解析器来解析网页和表格，而对于 PDF 和扫描图像，主流的做法是采用 OCR（Optical Character Recognition，光学字符识别）引擎来提取 VRDs 中的有效信息。OCR 通常包含两个模块文本检测模块和文本识别模块，分别来检测图片中文本块的位置并识别文本内容。OCR 识别的图片会得到两个重要的数据，即文本段和其边界框（bounding box）的位置坐标，每一个文本段和对应的边界框坐标被称为一个语义实体。因此，视觉富文档可以认为是由若干个语义实体在二维空间排版组成的。第二个阶段则基于这些语义实体构建文档智能模型来学习实体之间的联系，进而提升对 VRDs 的理解能力。

视觉富文档理解（VDU）包含多个下游任务，如文档图像分类，文档布局分析，文档信息抽取，文档视觉问答、文档布局生成等。本文的研究主要关注在文档信息抽取、文档视觉问答和文档布局生成这三个任务。

图 2.1　现有多模态视觉富文档理解的技术框架

进行视觉富文档理解的模型被称为视觉富文档智能模型（Visually Rich Doc-

ument AI model），在本文中简称为文档智能模型（Document AI Model）。文档智能模型发展到现在主要包含两种：多模态视觉富文档理解模型和大语言模型。近年来，大量的工作[3-5,8-9]集中在多模态视觉文档理解的研究，图 2.1 展示了基于多模态预训练模型的技术框架。从图中可以看出，针对不同的视觉文档理解任务，通常需要在特定的数据集上微调一个特定的模型，这严重影响了模型的领域迁移能力，模型学习的 VDU 知识是通用的，但因为任务的限制，不能有效的输出。而目前，大语言模型，如 ChatGPT、GPT4 等的相继出现，很好的利用通用提示模板的方式解决了统一 NLP 下游任务的问题，实现了进一步的通用化、智能化。图 2.2 展示的是基于大语言模型统一 VDU 下游任务的一般框架。

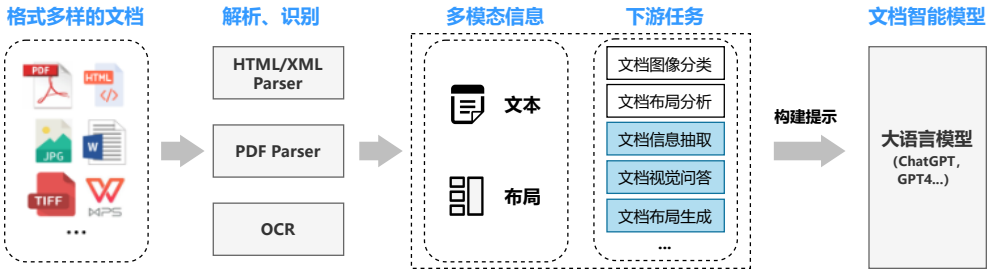


图 2.2　现有基于大语言模型的视觉富文档理解技术框架

2.2　多模态视觉富文档理解方法

2.2.1　光学字符识别 OCR

OCR（Optical Character Recognition，光学字符识别）技术能够将打印或手写的文本图像转换成可被编辑和处理的文本格式。OCR 在许多领域都有广泛的应用，包括：文档数字化、自动数据录入、票据识别、手写文字识别、文本检索等。随着深度学习和计算机视觉技术的发展，OCR 技术的准确性和应用范围不断提升，正在成为数字化转型和智能化应用中的关键技术之一。

OCR 技术主要涉及到两个模块：场景文本检测和文本识别。下面对这两个模块使用的技术进行概述：

（1）场景文本检测

场景文本检测的目标是检测自然场景（包括扫描图像）中的文本所在的区域，是文本识别的重要基础。由于场景文本的形状和排列方向可能非常多样，使

用传统的基于水平边界框的检测方法来自动处理这些文本存在一定的挑战。早期的文本检测方法主要是基于人工设计的特征表示，这导致检测的性能不佳。随着深度学习技术的兴起，基于深度神经网络的文本检测技术已经取得了显著的进展。下面对早期的文本检测方法和深度学习的方法分别进行介绍：

早期的场景文本检测方法：这些技术通常依赖区域特征，经典的方法有最大稳定极值区域[10]，文字戳宽度转换[11]。Pan 等人[12],Yin 等人[13]首先基于目标检测和感兴趣区域特征来检测字符级别的候选框，然后通过传统机器学习的分类模型对这些候选框进行链接并去除虚框，通过这种方式从图像中检测出单词或文本行边界。Yao 等人[14]也尝试探索了基于 SWT 特征的多角度文本检测方法。然而，由于传统的机器学习提取的特征和分类模型的性能通常无法达到预期，并且框架相对更加的繁琐，也涉及到很多人工的经验操作，逐步被基于深度学习的方法取代。更多早期的文本检测方法见 Ye 等人[15]的工作。

基于深度学习的文本检测方法：基于深度学习的文本检测方法大致可以分成三类：基于边界框（bounding box）回归的方法，基于图像语义分割（image semantic segmentation）的方法和融合回归与分割的方法。

基于边界框回归的文本检测方法需要模型能够准确地预测出文本区域的边界框顶点坐标。针对目前场景文本形状多变，排列方向自由和图像质量不佳的问题，CTPN[16]基于通用目标检测框架，将图像的像素特征空间分割为若干个子块，然后利用循环神经网络（RNN）学习字块之间的上下文联系。这种方法很好的处理了长文本检测的问题，但在非水平的文本上效果不佳。TextBoxes[17]和 TextBoxes++[18]通过设计多种长宽不同的边界框描点来适应文本的非水平变化，并同时预测水平和非水平边界框的位置。为了解决对任意方向的文本检测，Shi 等人[19]把文本的检测任务看成单词空间单元之间的关系预测，最后将单词空间单元根据先后关系组合得到任意方向的多边形，提高了非水平文本检测的鲁棒性。此后，ABCNet[20]引入贝塞尔曲线来拟合文本的形状，进一步提高了不规则文本的拟合精度，并且这种方法的高效使得模型可以在实时视频的场景下使用。此外，为了更贴近的拟合任意方向文本的轮廓，Zhu 等人[21]改用傅里叶曲线方法建模了文本轮廓。

基于图像语义分割的文本检测方法的目标是利用图像的像素级特征来适应

并识别出任意形状的文本。通常的语义分割方法把文本检测认为是前景和背景的像素分类任务，区分前景（通常是文本）和背景。这种方法依赖图像像素级别的预测概率分布，所以往往需要人工的后处理来得到更精确的检测结果。经典的方法如 FCN[22]，该方法第一次使用全卷积神经网络检测文本行的候选区域，并设计了基于区域特征抽取（MSER）设计了后处理方法，实现了多边形的文本检测。PixelLink[23]不仅仅关注于前景文本的像素级别预测，还考虑到文本像素之间的连接关系，这为后处理提供了更精确的文本区域。此后，典型的基于语义分割的方法大多关注在如何设计语义信息更准确的分割网络。例如，PSENet（Progressive Scale Expansion Network）[24]基于渐进式文本掩码的尺度拓展表示来优化目标分割网络，TextField[25]将文本掩码更换为以文本边界区分的方法像素场，进一步丰富分割网络的语义信息。CRAFT[26]以一种弱监督的方式引入字符的位置信息，这有利于文本区域分割的精准性。

融合回归与分割的文本检测方法主要是结合回归和分割方法的优势来精准定位文本所在的区域。典型的方法是 EAST（Efficient and Accurate Scene Text detector）[27]和 Direct Regression[28]，通过消除感兴趣区域的候选框的抽取，直接从全卷积分割图中预测任意文本四边形边框，这些方法不再依赖候选区域，取得了较好的检测结果。TextBPN[29]使用图神经网络迭代的优化由语义分割图抽取的粗粒度轮廓，从而得到了提高了检测的效果。近年来，基于 Transformer 的文本检测方法也取得了不错的成果。例如，Tang 等人[30]采用 Transformer 模型来构建和处理分割后的文本特征，这些特征随后被分成多个组，进而对每一组进行边界框坐标的回归分析。Song 等人[31]则利用视觉-语言预训练模型来攻克文本检测问题，实验表明该预训练模型在两种不同的检测方法中均能显著提高性能。

（2）文本识别

文本识别旨在识别图片中文本行的内容，因为文档中的行列排布整齐，比较容易识别，而文本段中的字符切分却十分有挑战。文本识别根据图像内容的不同通常包括手写文本识别和场景文本识别两个方向。本文主要关注视觉富文档场景的文本识别，手写文本识别亦不是本文的重点。因此，下面主要介绍场景文本识别的主要方法。

场景文本识别：早期研究中，主要采用的是对字符进行分割的技术，一个典

型的例子是谷歌在 2013 年提出的 PhotoOCR[32]。这种方法通过将文本字符进行分割，替代了传统的字符分类器，从而简化了识别器的设计过程。然而，这也增强了后处理阶段对字符分割准确性的依赖性。因此，现阶段大部分的研究都基于端到端（end-to-end）的方式来处理文本识别任务。基于端到端的文本行识别方法通常被划分成基于 CTC 的方法和基于注意力机制（attention mechanism）的方法。CTC 方法旨在解决输入序列和输入序列长度不一，难以对齐的问题。它提出了一种新的对齐方式, 将输出序列切分为多个时间段, 通过制定一定的处理规则来实现序列对齐。Shi 等人[33]提出 CRNN 模型，巧妙的将文本识别任务转化为序列标注（sequence annotation）的任务，该方法可以适应多种文本识别的场景，取得了不错的成果。Yin 等人[34]在此基础之上提出一种基于滑动窗口分类的文本识别方法，有效地解决了 RNN 本身梯度消失、收敛速度慢的问题。

基于注意力机制的方法最早由 Shi 等人[35]引入文本识别领域，由于注意力机制强大的能力，该方法迅速成为文本识别领域的重要研究内容。注意力机制解决了文本识别任务中的诸多难题，包括 Li 等人[36]将一维注意力推广到更适合图像特征的二维注意力，Cheng 等人[37],Wang 等人[38]解决了注意力焦点偷换（注意力漂移）的问题，针对不规则场景文本提出的矫正识别方法，如 ASTER[39]和 MORAN[40]，基于 Transformer 架构的预训练语言模型的方法[41]。最近，随着视觉的 Transformer（ViT）的出现与研究，基于 ViT 设计的一系列文本识别方法[42-43]开始涌现，在这个领域取得了突破性的成果。

端到端的文本检测与识别：随着文本识别技术的进步，越来越多的研究将图像中的文本检测和识别视为一个连贯的端到端流程，即同时处理检测和识别任务。目前，基于深度学习的端到端文本检测与识别方法大致有两个方向：两阶段方法和单阶段方法。在早期的两阶段方法中，核心策略是让文本检测组件和识别模型共用同一主干网络来提取特征，接着利用这些特征在检测组件中定位文本区域，最终将这些区域的特征传递给识别模块以完成文本的识别工作。Li 等人[44]提出了一个经典的端到端两阶段方法，但该方法难以处理非水平的文本。随后，Liu 等人[45]提出了对无规则文本四边形的端到端改进模型。近年来，两阶段的研究更多关注在非水平文本的检测与识别，例如 Feng 等人[46]将字符和文本块看作是组件，从而更灵活的进行组合来解决非水平文本的问题；ABCNet

一族[20]的核心想法是利用贝塞尔曲线或傅里叶曲线建模非水平文本的几何边界；Mask TextSpotter 一族[47-48]的方法融合了前述两种技术的优势，它采用 Mask R-CNN 来定位文本行，并结合字符检测机制，以更灵活的方式识别非水平方向的文本。上述的两阶段方法，仅仅只是共享检测模块和识别模型的视觉特征，而 Huang 等人[49]提出一种让识别模块可以接受检测模块信息反馈的模型，大大提高了文本识别的效果。

然而，这种两阶段的端到端方法在进行特征转换和对齐的时候难免会导致信息的损失，进而影响文本识别的精度。近年来，也有许多工作聚焦在单阶段的研究。单阶段方法的重点是让模型可以在检测出文本区域特征的同时，识别区域内的文本特征。经典的方法如 CharNet[50]，通过文本检测时的单字检测来识别文本行的内容；MANGO（Mask AttentioN Guided One-stage）[51]利用文本检测阶段的字符分类结果，输出文本内容的识别结果，并实现了并行化；Zhang 等人[52]提出 TESTR，该方法根据 DETR[53]设计了一种双子解码器，可以同时检测文本边界框并识别框中的内容；SPTS[54]首次以序列建模的角度看待文本检测和识别任务，并采用基于 Transformer 的序列预测模型进行端到端的检测与识别。目前，基于 Transformer 框架的单阶段方法正在成为端到端文本检测与识别的主流方法。

2.2.2 版面布局分析

文档版面分析（Document Layout Analysis，DLA）的目的是利用文档区域之间的语义差异，将文档图像划分为多种不同的区域类别，包括文本段落、图片、表格和公式等，并进一步探究这些区域间的语义联系。版面分析任务通常被分为两大类：针对页面区域分割的称为几何版面分析（物理），针对区域之间语义关系和人类阅读顺序的被成为逻辑版面分析。下面分别介绍这两类任务的经典方法：

（1）几何版面分析

早期的几何版面分析主要关注印刷或扫面文档的物理分割。根据分割角度的不同，页面物理分割的方法大致可以分为三种：自上而下的页面分割方法、自下而上的页面分割方法和融合两种方式的页面分割方法。自上而下的方法依赖于文档的整体信息，通过执行直方图投影或背景迭代等操作来进行分割，从而渐

进式的将文档中大的区域分割称较小的区域。典型的方法是 X-Y Cut 算法、背景空白细化等。自下而上的方法利用底层元素（像素点和连通分量等）进行聚类的方式将较小的区域逐步合并成较大的区域，这种方法的代表性例子包括游程平滑[55]、Voronoi[56]等。混合方法则是将上述两种方法结合，利用文档全局的先验信息迭代地聚合或分裂区域。相对来说，自下而上的方法和融合的方法不会过于关注复杂文档的整体结构，因此更适合版面结构复杂的文档图像。

随着深度学习技术的不断发展，基于神经网络的版面分析方法开始涌现。和传统算法相比，神经网络能从数据中学习更有效的特征并进行不断优化。目前，基于深度学习的几何版面分析方法主要分为目标检测和实例分割方法。基于目标检测的方法[57]使用计算机视觉（Computer Vision，CV）中的通用目标检测器检测文档图像中的感兴趣区域，如文本段，图片，表格和公式等，其中典型的通用目标检测器包括 Faster R-CNN[58]、Mask R-CNN、YOLO 一族[59]的模型。这类方法在识别尺寸适中的矩形区域方面表现出色，但对于任意形状，尤其是弯曲或倾斜的区域时，它们不能很好地处理。为此，基于实例分割的方法能很好的分割任意形状的区域。这种方法主要是利用全卷积网络（FCN）或 Mask R-CNN 等实例分割框架来进行页面分割，例如分割文本段区域，Yang 等人[60]更多的考虑分割类型为表格、图片和公式的区域。此外，随着图神经网络的发展，关于图模型的分割方法在版面分析领域大量出现，这种方法的核心是将自底向上抽取的单元区域通过图卷积网络（GCN）[61]或图注意网络（GAT）[62]聚合成目标区域。图神经网络的方法结合了自上而下的全局先验和自下而上的单元聚合，从而适合处理版面复杂的文档图像，但这种方法需要额外使用目标检测算法或者人工的方法先得到基本的单元区域。

（2）逻辑版面分析

逻辑版面分析主要关注在文档图像中文本区域的语义类型，比如标题、摘要、正文等，并对这些区域文本的阅读顺序进行分析。早期的逻辑版面分析方法主要基于启发式规则，围绕文本内容、文本位置和一些文字级别的细粒度特征进行文本区域的语义分类。传统的文档人类阅读顺序的分析着重于利用领域文档版式结构或人类阅读习惯（如从左到右，自上而下），设计一系列规则来判定文档的阅读顺序，这种方法往往只能处理版式结构固定、阅读顺序常规的文档，缺

乏领域的泛用性和拓展性。随着深度学习的发展，基于神经网络的阅读顺序模型[8,63]，利用文档中丰富的视觉特征、布局位置和文本信息对人类阅读的顺序进行了精准的预测。

本文关注的视觉富文档不仅包含普通的文本内容，往往还存在结构化和半结构化的表格。近年来，表格结构的解析是逻辑版面分析的热门研究方向。本文把这些工作大致分为三类：自上而下的表格结构解析方法，自下而上的方法和基于表格到标记（table-to-markup）的方法。自上而下的方法从表格的全局信息出发，先通过页面分割的方式通过页面分割技术提取出表格内的行与列，继而通过计算这些行列的交汇部分来确定单元格的具体位置。显而易见，该方法主要适用于行列规整的结构化表格，对于表格结构松散，内容稀疏，存在跨行跨列的情况则效果不佳。自下而上的方法[64-65]则从文档底层的元素出发，使用 OCR 或目标检测的方式抽取文本单元，再通过卷积、Transformer 等模型把具有相似语义和位置的文本单元聚合成一个单元格，同时预测单元格之间的位置关系。和自上而下的方法对比，自下而上的方法更适合处理表格结构复杂（如跨行跨列、无边框线）的场景，但该方法的性能受限于 OCR 和检测算法的精度。基于表格到标记的方法[66]把表格的表示看作是形式多样的标记语言，如 HTML、XML 和 Latex 等，该方法主要利用了编码器-解码器（encoder-decoder）的结构，效仿图生文（image captioning）的方式，用卷积神经网络编码特征，然后用注意力机制解码，直接从表格图像生成表格的 Latex 等的源码。此方法的优点是训练数据庞大，因为从表格的 Excel 或 PDF 形式中获取标记源码的监督信息十分便捷；但该方法也有明显的局限性：无法直接给出表格行列、单元格的位置信息，表格结构难以复现，模型可解释性的问题。

2.2.3 多模态表示学习

本小节主要介绍本文关注的文本、图像特征表示的发展历程，文本和图像的特征在视觉富文档的理解中具有重要的地位。下面从基于预训练语言模型的文本特征表示和基于卷积神经网络的图像特征表示分别进行概述。

（1）基于预训练语言模型的文本特征表示

随着深度学习的快速发展，自然语言处理的文本表示大致经历了三个发展

阶段：基于预训练词嵌入（word embedding）的表示方法、基于预训练上下文编码器的表示方法和基于大语言模型的表示方法。如图 2.3 所示。

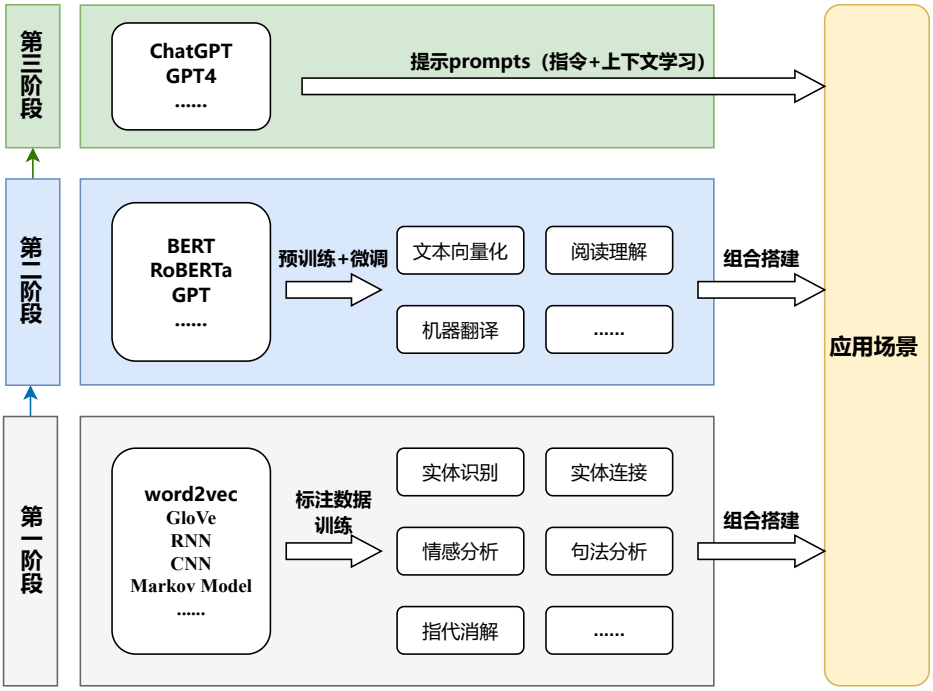


图 2.3 预训练语言模型发展的三个阶段

在自然语言处理（NLP）的早期发展阶段，单词通常被编码为紧凑的独热（one-hot）向量。随后，神经网络语言模型（NNLM）[67]成为了引入词嵌入技术的关键。Collobert 等人[68]证实，通过在非标注数据上进行预训练得到的词嵌入能够有效提升多种 NLP 任务的表现。为了解决计算复杂性，他们将语言建模替换为成对排名任务来学习词嵌入。Mikolov 等人[69]进一步指出，创建有效的词嵌入并不必然依赖于复杂的深度神经网络结构。他们提出了两种高效的模型架构：连续词袋（CBOW）和跳词（Skip-Gram）模型，这些模型不仅易于实现，还能捕捉单词间的潜在语法和语义相似性，从而生成高质量的词嵌入。Word2Vec[70]便是这种技术的一个典型实例。同样，GloVe[71]也是一种广泛使用的方法，它通过分析大规模文本语料库中的全局词-词共现统计信息来生成预训练的词嵌入。虽然基于预训练词嵌入的方法在许多 NLP 任务取得了一定程度的效果，但这些方法是上下文无关的，并且在任务迁移时需要从头开始训练。

在 NLP 的多个应用中，理解句子乃至整个段落的上下文对完成任务至关重

要，这不仅仅是对单个词汇含义的关注。基于此，对神经网络编码器在句子或篇章层面进行预训练是合理的。这些编码器生成的输出向量被称作上下文词嵌入，因为它们能够反映出单词在特定语境中的语义。Dai 等人[72]利用语言模型（LM）或序列自编码器对长短期记忆网络（LSTMs）进行预启动，观察到这一步骤能够增强 LSTMs 在文本分类任务中的学习和泛化能力。ULMFiT（Universal Language Model Fine-tuning）是一种在文本分类任务中对预训练语言模型进行微调的方法，它在多个文本分类数据集上实现了突破性成果。ULMFiT 包括 3 个阶段：1）在通用领域数据集上对语言模型进行预训练（学习通用知识）；2）在领域数据集上进行微调（学习领域知识）；3）在目标任务上进行微调（适应特定任务）。

近年来，基于大规模数据集预训练的语言模型在通用语言表示能力上展现了巨大的潜力，例如 BERT（Bidirectional Encoder Representation from Transformer）和 OpenAI 的 GPT（Generative Pre-training）。由于这两个模型是本文研究的基石工作，所以下面详细介绍这两个模型的区别：

BERT[73]（Bidirectional Encoder Representations from Transformers）是一种由 Google AI Language 团队于 2018 年提出的预训练语言模型。BERT 模型基于 Transformer 架构，该架构由 Google 在 2017 年提出，用于处理序列到序列的自然语言处理任务。与传统的从左到右或从右到左的语言模型不同，BERT 是一个双向的模型，可以同时考虑一个单词左右两侧的上下文信息。BERT 的预训练过程采用了两种新颖的技术：掩码语言模型（Masked Language Model，MLM）和预测下一句（Next Sentence Prediction，NSP）。在 MLM 任务中，输入序列中的一部分单词会被随机遮盖，模型需要预测被遮盖的单词是什么。在 NSP 任务中，模型需要判断两个句子在语义上是否连续。通过这两种任务的预训练，BERT 可以学习到更深层次的语言表示。在预训练完成后，BERT 可以通过微调来适应各种自然语言处理理解任务，例如文本分类、命名实体识别、问答等。各项研究表明，BERT 在多个自然语言处理任务上取得了最先进的性能，为后续研究带来了深远的影响。

GPT[74]（Generative Pre-trained Transformer）是由 OpenAI 于 2018 年提出的生成式预训练语言模型。GPT 模型采用了 Transformer 架构，它通过大规模的无监督学习在文本数据上进行预训练，然后可以在各种自然语言处理任务上进行

微调。GPT 模型的核心思想是使用自回归（auto-regressive）的方式进行预训练。在预训练阶段，模型通过观察前面的文本来预测下一个单词或标记的概率分布。这种方法使得模型能够理解语言的结构和语境，并学习到丰富的语言表示。与 BERT 的掩码语言模型 MLM（自编码模型）不同，自回归的模型更擅长生成式 NLP 任务，如对话生成、语言翻译和摘要生成等，并且可以简单地将训练目标设置为预测语料库中的下一个单词，因此生成数据相对容易。但这种自回归的方式只能前向或者后向建模，无法同时捕捉上下文单词的内在联系。

最近，随着预训练数据集和模型参数量的急剧膨胀，出现了以 GPT3[75] 为首的超大规模预训练语言模型，简称为大语言模型。与之前的预训练语言模型，如 BERT，不同，GPT3 拥有 1750 亿的参数量，是 BERT 模型的成百上千倍。在如此巨大的参数规模下，GPT3 模型中蕴含了庞大的知识，具有十分强大的语言理解能力，但这也导致微调模型的代价急剧的增大。因此，上述的预训练-微调的范式已经不再适合大语言模型，随着指令微调[76] 和上下文学习[77] 的研究逐渐成熟，新的基于大语言模型的提示方法成为最新的范式。关于大语言模型更详细的介绍请参照第 2.3 节。

（2）基于卷积神经网络的图像特征表示

卷积神经网络（CNN）是一种深度学习模型，在图像处理领域取得了巨大的成功[78]。CNN 的核心思想是通过卷积操作来提取特征并学习表示数据。它包括多个卷积层、池化层和全连接层组成的堆叠。其中，卷积层通过对输入数据进行卷积操作，从而提取输入数据的局部特征。池化层用于减少特征图的尺寸，同时保留最重要的信息。全连接层用于将卷积层提取的特征映射到最终的输出类别或标签。自从 CNN 出现后，图像领域大量的工作集中在优化 CNN 网络结构和减小模型复杂度，目的是为了使用更小参数量的模型解决更复杂的任务。目前，典型的基于 CNN 的特征表示模型有 AlexNet、VGG-Nets、ResNet 等。下面分别进行简要的介绍。

AlexNet[79] 是对 CNN 深刻理解并应用的网络模型，它包含了 8 层神经网络，其中包括 5 个卷积层和 3 个全连接层。该模型第一次采用了 ReLU 激活函数以及 Dropout 技术。此外，AlexNet 还首次使用了 GPU 来加速深度学习的训练过程，极大地提高了训练速度。其解决了模型过拟合和梯度消失的问题。赵等人[80] 基

于 AlexNet 提出一种优化方法，减少全连接层并引入 SE-Block 模型加快模型训练速度，但此方法对于生物特征，如断掌的识别效果不佳。

VGG-Nets[81]模型是由牛津大学与 Google DeepMind 公司共同联合研究开发的一种神经网络结构。该模型通过堆叠卷积层和池化层构建了一种深度可达数十层的卷积神经网络。这种深度的结构使其能够更好地捕获图像中的特征，并在各种视觉任务中取得显著的性能。为了简化网络结构，VGG-Nets 模型采用了小尺寸的卷积核（1×1 和 3×3），并使用了 Softmax 逻辑回归作为分类器。尽管这种设计增强了网络的表示能力，但也增加了计算量，并且因为网络变得更深，特征的传递变得更加困难。

ResNet[82]是卷积神经网络（CNN）领域的一项创新，其独特之处在于引入了残差模块进而构建一个非常深的网络结构。相较于传统的网络结构，ResNet 模型大幅减少了参数量和计算量，同时解决了网络退化的问题，这使得神经网络在训练过程中能够更加高效地学习到数据的特征。这种创新的关键在于残差连接，残差连接让神经网络仅需学习输入与输出之间的差值，从而简化了问题的复杂性。这一举措有效地解决了以往在深度网络中出现的信息丢失、梯度消失以及神经网络无法正常训练等问题。通过保留残差信息，网络可以更深地传播梯度，从而更好地进行反向传播，进而让网络的训练变得更加容易和高效。Tao 等人[83]在这一基础上进一步提出了一种改进方法，他们采用 ResNet-50 网络作为基础，将两个卷积层替代全连接层，并使用 Soft-NMS 算法来增强鲁棒性，提高了检测精度。总之，ResNet 相较于之前的方法在速度和精度方面都取得了显著的提升，展现出了极大的潜力和应用前景。

2.2.4 版式感知的文档理解模型

Transformer：Transformer[84]是由 Google 在 2017 年提出的一种用于序列到序列（sequence-to-sequence）学习的模型架构，它在自然语言处理领域取得了巨大的成功。Transformer 的设计主要是为了解决长序列数据处理中的效率和性能问题，尤其是在机器翻译等任务中，传统的循环神经网络（RNN）和长短期记忆网络（LSTM）存在着难以并行化和难以捕捉长距离依赖的问题。Transformer 模型由编码器（encoder）和解码器（decoder）组成，其中编码器用于将输入序列

转换为隐含表示，解码器用于根据编码器的输出生成目标序列。

Transformer 模型的核心思想是利用堆叠的自注意力机制（Self-Attention Mechanism）来建立序列之间的依赖关系。自注意力机制是注意力机制的一种。在注意力机制中，每个输入位置都会分配一个权重，该权重表示该位置对当前位置的重要性。这些权重是通过计算当前位置与所有其他位置之间的相似度得到的，并经过 softmax 函数归一化后得到。然后，模型使用这些权重来对其他位置的表示进行加权求和得到当前位置的表示。这种机制的优势在于（1）并行计算：自注意力机制允许模型并行地计算序列中每个位置的表示，提高了模型的训练和推理效率。（2）长距离依赖：自注意力机制能够在不损失信息的情况下捕捉长距离的依赖关系，使得模型能够更好地理解较长序列数据。（3）灵活性：自注意力机制具有很高的灵活性，可以根据任务和数据自适应地学习到不同的依赖关系。

现有版式感知的富文档理解方法：在文档理解领域，基于规则的方法[85]是重要的。这些方法通过采用特定配置规则为每个模板设计，解决文档理解问题。然而，基于人工规则的模板匹配方法[86]需要为不同类型、不同版式结构的文档设计不同的规则，这往往会耗费大量时间和精力。另一方面，基于文字流和分布式环境下的表格信息抽取方法仅适用于特定类型的表格和文档，需要预先制定特定的模板或关键词信息。然而，基于模板或文字流的表格信息抽取方法在面对不存在完整表格的表单时效果较差，通用性也较低。这些挑战表明，尽管规则方法在文档理解中发挥着重要作用，但仍需要更灵活和智能的方法来解决信息抽取的复杂性和多样性问题。

随着深度学习的快速发展，许多基于深度学习的视觉富文档信息抽取（VIE）方法已经出现[87]，这些方法不仅在准确性和能力方面显著优于传统的基于规则和基于模板的方法，而且在处理复杂文档结构和多模态信息方面表现出了优势。VIE 方法将文档抽取视为一个 token 分类问题，即将文档中的每个 token（例如单词、数字、符号等）与其所属的字段类型进行关联。这些方法使用不同的深度学习模型，例如 CNN、RNN 和 Transformer 等，来预测每个文档 token 的字段类型。早期的工作[88]主要采用语言模型从纯文本中提取实体信息，然而，随着研究的不断深入，越来越多的工作开始关注布局信息与文档表示的结合。这些研究

通过构建 token 级别的二维网格模型，旨在挖掘文档中的上下文嵌入信息。在文本信息的基础上，一些学者也开始认识到整合多模态信息的重要性，并着手通过结合视觉和布局信息来增强 VIE 的性能与鲁棒性。主要采用的方法包括图神经网络和 Transformer 等模型，这些模型有助于更有效地整合不同模态的信息，在实体抽取等任务中取得了最先进的结果。根据不同的文档表示，VIE 方法大致可以分为三种：基于网格、基于图和基于序列。

基于网格的方法，如 Chargrid[89]，将文档视为由 token 嵌入构成的二维网格，并利用实例分割模型来提取字段值。这些方法的优势在于能够同时考虑文本和布局特征，但却忽视了视觉特征的重要性。为了解决这一问题，VisualWordGrid[90]方法将网格表示与文档图像特征相结合，创造了更为综合的多模态二维文档表示。这种方法将文本、视觉和布局信息相结合，提升了文档理解的性能和鲁棒性。然而，与当前最先进的技术（如 LayoutLM、PICK）相比，基于网格的方法存在两个主要缺陷：首先，它们未能充分利用最先进的上下文词嵌入技术来生成强大的 token 嵌入；其次，它们在训练过程中使用了固定的 BERT 预训练参数，没有充分挖掘 token 嵌入的潜力。为了解决这些问题，最近的研究提出了 ViBERTgrid[91]方法，通过联合训练 BERT 和 CNN 参数，显著改进了文档理解的效果，并在多个数据集上取得了显著的提升。

基于图的方法在文档理解领域中扮演着重要角色，它们将文档构建成一个图，其中每个边界框（bounding box）对应于图中的一个节点。这些节点的初始特征可以基于单一的信息模态，也可以是文本、视觉、布局等多种模态特征的融合。通过应用图神经网络或自注意力机制，能够对图中节点的特征进行更新，以更有效地捕捉文档的结构和内容。在一些方法中[92-94]，图中的每个节点嵌入到相应文本段中的所有 token 嵌入，然后输入序列标注模型提取字段值。最近，像 PICK[93]这样的基于图的方法通过图卷积技术，充分利用了文档的多种多模态特征，在公共的 SROIE[95]数据集上展现了卓越的性能。这些方法不仅提高了文档理解的准确性，还为处理复杂文档结构和多模态信息提供了一种有效的方法。

基于序列的方法首先将文档图像序列化为一维文本序列，然后利用现有的 NLP 序列标注模型[73,96-97]来提取字段值。早期的方法[98-100]试图在 token 嵌入中引入二维位置编码，但仍然依赖于文本段序列化的准确性，这使得它们难以应

用于复杂布局的文档序列化。为了解决这些问题，现有的方法利用预训练语言模型，如 BERT，来学习有效的上下文语义表示。例如，Post-OCR[100]通过微调 BERT 来解决 VIE 任务。另一个典型的方法是 BERTgrid[101]，它在图像分割任务中使用 BERT 将文本信息编码到图像像素中。然而，由于 BERT 是通过纯文本预训练的，将文本信息转换为单一的文本序列会导致空间布局信息的丢失，这可能限制了模型对文档整体结构的理解。

（1）**LayoutLM**[5]：微软亚洲研究院提出了 LayoutLM 模型，该模型利用文本块的空间信息在大规模文档数据上进行预训练。LayoutLM 通过考虑文本块在页面布局中的位置和结构等信息，在视觉富文档理解的多个下游任务上展现出了较好的性能，这一结果证明了预训练方法在文档理解中的有效性。相较于传统的文本预训练模型，LayoutLM 的优势在于它不仅处理了文本信息，还深入挖掘了文本在页面版式中的位置信息，这些位置信息增强了模型对文档内容的识别与理解能力。LayoutLM 模型已经在各种文档理解任务中得到广泛应用，其中包括但不限于文档分类、信息抽取、表格识别和布局分析等领域。其在多个任务上的成功应用表明了该模型在处理不同类型的文档和理解复杂文档结构方面的优越性，为文档理解领域的进一步发展提供了重要的参考和启示。

LayoutLM 是 BERT 模型的二维版本，主要的区别是引入了重要的布局特征，即二维位置编码。此外，LayoutLM 采用了两种不同的预训练任务：掩码视觉语言模型（MVLM）和多标签文档分类（MDC）。通过这两个预训练任务，LayoutLM 能够在文档理解任务中更好地利用位置信息和图像特征，为处理视觉富文档和实现多模态信息融合提供了强大的基础。

然而，LayoutLM 模型有三点局限性：第一，LayoutLM 引入了文本边界框的位置信息，但它并没有充分考虑文本序列中的位置与实际二维空间之间的差异。。这导致模型无法充分利用空间信息，并可能在处理特定类型的文档时表现不佳。第二，LayoutLM 的预训练目标中没有明确考虑文本块之间的空间关系。这意味着 LayoutLM 可能未能充分利用文本块之间的空间信息，限制了其在实际任务中的性能。这些限制表明 LayoutLM 在处理文档理解任务时可能无法充分利用空间信息。

为了解决这些问题，LayoutLMv2 提出空间感知的自注意力机制，在输入时

整合图像编码，构建布局感知的多模态预训练模型，从而克服 LayoutLM 的局限性，提高模型在复杂文档理解任务中的性能和鲁棒性。

（2）**LayoutLMv2**[3]：LayoutLMv2 标志着多模态预训练模型领域的一个显著进步，它整合了文本、图像以及布局信息。与前身 LayoutLM 模型相比，LayoutLMv2 的关键创新在于它不仅纳入了图像信息，还采用了能够感知空间关系自注意力机制。在处理输入时，首先将图像均匀划分为多个图像块（patch），再利用 ResXNet 抽取视觉特征，最后通过注意力机制与文本、空间信息融合。此外，模型在一维位置编码的基础上加入了二维位置编码，旨在捕捉不同文本块之间的相对空间关系，以更准确地捕捉文档的结构和布局。为了进一步提高文档理解的准确性，LayoutLMv2 模型采用了两种新的预训练任务—“文本—图像对齐”（Text-Image Alignment）和“文本—图像匹配（Text-Image Match）”。通过这两个任务的联合训练，模型可以高效地学习文本和图像之间的跨模态对齐，为文档理解任务提供更加全面和准确的处理能力。

（2）**LayoutLMv3**[4]：LayoutLMv3 是微软亚洲研究院为了克服文本和图像在预训练任务上的差异而提出的改进模型，其主要目标是统一文本和图像的预训练过程，以更好地理解和处理视觉富文档。为了实现这一目标，LayoutLMv3 采用了统一的文本和图像掩码目标，即掩码语言模型（Masked Language Modeling，MLM）和掩码图像模型（Masked Image Modeling，MIM），用于对文档模型进行多模态 Transformer 预训练。在 LayoutLMv3 中，模型通过学习重建文本和图像模态的掩码标记来实现跨模态对齐，并引入了单词图像块对齐（Word-Patch Alignment，WPA）目标，以预测文本单词的对应图像块是否被掩码。这样的设计使得模型能够更好地理解文本与图像之间的关联性，从而提高对视觉富文档的理解和处理能力。此外，受到 ViT 和 ViLT 启发，LayoutLMv3 直接利用文档图像中的原始图像块，无需进行复杂的预处理步骤，如目标检测。值得注意的是，LayoutLMv3 采用了统一的 MLM、MIM 和 WPA 目标，以联合学习图像、文本和多模态表示，使其成为首个不使用图像嵌入的多模态预训练文档理解模型。这一设计不仅节省了参数空间，还提高了模型的通用性，使其能够适用于图文混排的理解任务。

2.3 基于大语言模型的视觉富文档理解方法

本节主要介绍基于大语言模型的视觉富文档理解方法，首先从预训练、微调、推理和评估四个角度概述大语言模型的通用技术，然后介绍在视觉富文档领域内大语言模型的应用方法。

2.3.1 大语言模型概述

（1）**大语言模型的发展**：GPT 系列模型是大语言模型发展的重要里程碑。GPT-1[74]，作为该系列的第一款模型，由 OpenAI 于 2018 年发布。它采用了生成式的、仅含有解码器的 Transformer 架构，并采用了无监督预训练和有监督微调相结合的方法。这一核心架构奠定了后续 GPT 系列模型的基础。随后在 2019 年发布的 GPT-2[102]，规模达到了 15 亿参数，使用了大规模的网页数据集 WebText 进行训练。其目标是通过无监督自回归建模来执行任务，而无需大量的标注数据，从而扩展了 GPT 系列模型的适用范围。GPT-3[75] 作为 GPT 系列的下一代，于 2020 年发布，参数规模扩大至 1750 亿，并引入了上下文学习[77]（In-Context Learning，ICL）的概念。这一模型不仅在各种 NLP 任务中表现出色，而且在需要推理或领域适配能力的任务中也表现出色。GPT-3 可以被视为从预训练模型到大语言模型进化过程中的一个重要里程碑。它证明了将神经网络扩展到大的规模可以大幅增加模型的能力。

为了开发更强大的大语言模型，OpenAI 探索了两种方法来进一步改进 GPT-3 模型，即代码训练与人类反馈的强化学习。承载着这两种机制，ChatGPT 于 2022 年 11 月发布，专注于对话能力的优化。该模型的训练数据通过将人类生成的对话与 InstructGPT[76] 数据集结合而生成，同时支持了插件机制，进一步扩展了其功能，为对话系统领域带来了新的发展。

最新的 GPT-4 于 2023 年 3 月发布，它不仅将处理的文本输入扩展到了多模态输入，还在处理复杂任务方面展现了显著增强的能力。同时，GPT-4 对于具有恶意或挑衅性的提问的响应更加安全，采用了多种干预策略来减轻语言模型的潜在问题，如幻觉、隐私和过度依赖。GPT-4 是目前最先进的大语言模型之一，它具备的强大能力将深刻改变这个世界。近几年主要的通用大语言模型的发展如图 2.4 所示。

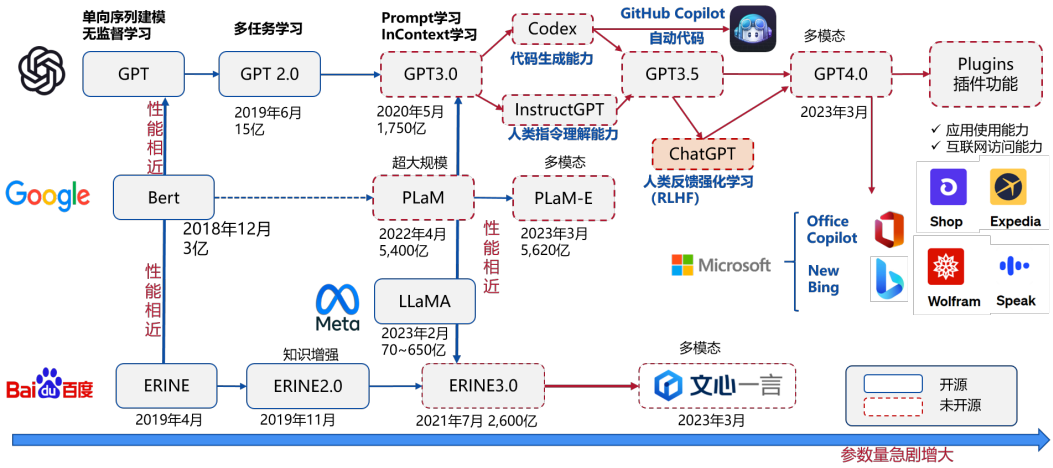


图 2.4　通用大语言模型的发展时间轴

（2）**预训练方式**：大语言模型的预训练通常分为三个阶段，无监督训练、有监督训练和人类反馈的强化学习。这三个阶段逐步的将大量知识注入到模型中，并进行约束和人类偏好的对齐。大语言模型的三阶段预训练如下所示。

- **无监督训练阶段**：大语言模型的成功离不开对大量数据的收集、清洗以及模型的构建和预训练过程。首先，模型的预训练数据来自于网络上的对话、社交媒体平台的对话记录以及电子邮件等多种来源。这些数据用于训练模型，使其能够学习到广泛的对话主题和语言风格。然而，从不同来源收集的数据可能存在噪声或无用信息，因此在数据清洗阶段需要对数据进行处理，以确保模型在训练过程中学习到高质量的对话内容。接着，大语言模型通常都采用 Transformer 模型作为核心架构，包括编码器和解码器部分，用于处理输入文本并生成输出文本。在模型构建完成后，进行无监督预训练。预训练过程通常采用自回归方式，即模型根据之前生成的单词预测下一个单词。在预训练过程中，模型只需要原始的文本作为输入，无需人工标注的标签，这使得预训练过程更加高效和灵活。通过无监督训练的阶段，大语言模型可以从数据中学习海量的知识，为后续的对话任务奠定了坚实的基础。
- **有监督训练阶段**：有监督训练是为了约束大语言模型在上个阶段学习的海量知识，避免大语言模型生成一些有毒有害的内容，这个阶段的预训练是为了调整模型的输出规范。有监督训练过程通常需要一些带标签的对话数据，用于调整模型的对话参数，使其更好地适应特定的对话风格。微调过

程通常比较快速，因为模型已经在无监督预训练阶段学习到了大量的语言知识和对话模式。此外，在此阶段预训练完成之后，模型会产生理解指令要求和范例要求的能力，这是上下文语境学习的基础。

- **人类反馈强化学习阶段（RLHF）**：RLHF 是为了解决有监督训练阶段造成的模型输出过于模板化、创造受限的问题。为了使大语言模型与人类的价值偏好对齐，研究人员通过收集到的人类反馈数据对大语言模型进行进一步的训练。RLHF 采用强化学习算法，例如近端策略优化[103]（Proximal Policy Optimization, PPO），通过学习奖励模型来优化大语言模型的输出。经过此阶段的学习，大语言模型的输出会更符合人类的认知和价值观，在语言上具有一定的创意，进一步防止模型输出有毒有害的内容。

（3）**微调方法**：与传统的预训练模型的全量微调方式不同，巨大规模的参数量使大语言模型进行一次训练的成本十分昂贵，所以低成本的微调方法成为目前大模型适应特定任务的主流方法。轻量级的微调方法主要包括 adapter tuning、prefix-tuning、p-tuning v2[104]和 LoRA[105]。然而，除了 LoRA 外，其他三种微调方法都有严重的局限性，如 adapter tuning 增加了模型的层数，这会增加模型的推演延迟；prefix-tuning 训练困难，并且 prompt 序列占用了下游任务的序列空间，导致模型性能受限；而 p-tuning v2 有灾难性遗忘的问题，微调后的模型，在之前的通用任务上的表现显著下降。

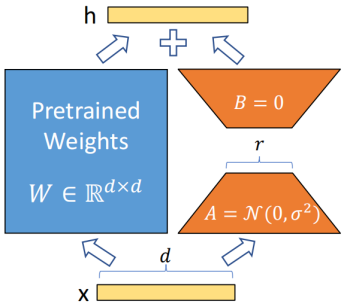


图 2.5　LoRA 低秩优化的原理[105]

LoRA 是一种基于前人有关模型内在维度研究的背景下提出的方法，该研究指出模型是过参数化的，通常具有较小的内在维度，并且模型主要依赖于这个低维度来适应特定任务。LoRA 方法建立在以下假设和方法基础上：首先是低秩假设，即假设在模型适应任务的过程中，权重的改变量是低秩的。基于这一假设，

提出了低秩自适应（LoRA）方法。LoRA 方法通过在神经网络的某些密集层进行训练时，采用秩分解的方式对这些层的权重矩阵进行优化，从而间接地调整网络的适应性，而在此过程中，预训练的权重保持不变。LoRA 方法的核心思想如图 2.5 所示。首先，在原始的大语言模型旁边增加一个旁路，进行降维再升维的操作，以模拟内在秩。其次，通过随机高斯分布初始化 A，用零矩阵初始化 B，以确保在训练开始时，此旁路矩阵为零矩阵。然后，在训练过程中，固定模型的参数 W，只训练降维矩阵 A 和升维矩阵 B。模型的输入输出维度保持不变，输出时将 BA 与 W 的参数叠加。这种思想类似于残差连接，并使用旁路的更新来模拟全量微调的过程。

LoRA 方法通过在神经网络的某些密集层进行训练时，采用秩分解的方式对这些层的权重矩阵进行优化，从而间接地调整网络的适应性，而在此过程中，预训练获得的权重保持原状不变。

（4）**推理**：推理时的涌现能力是区别大语言模型和普通的预训练模型的最显著特征之一。大语言模型最典型的三种涌现能力为上下文语境学习（ICL）、指令遵循和思维链（Chain-of-Thought, CoT）推理。以下简要介绍这三种能力：

- **上下文语境学习**[77]：上下文语境学习是指 GPT-3 引入的范例学习能力，它赋予模型在未经额外训练或梯度更新的情况下，根据输入文本的单词序列生成预期的输出的能力。例如，为大语言模型提供几个特定任务的范例，大语言模型能容易的迁移到这个特定任务的知识范围进行回答，这种能力同时也依赖于大语言模型强大的零样本和小样本学习的能力。GPT-3 模型在一般情况下就会展现出强大的 ICL 能力，而 GPT-1 和 GPT-2 则缺乏这种能力。
- **指令遵循**[76]：指令遵循是指通过指令微调，大语言模型在未见过的任务指令上表现出色的能力。大语言模型能够在没有显式示例的情况下遵循新的任务指令，因此具有更好的泛化能力。大量研究发现[106-107]指令微调在模型大小达到 680 亿时表现良好，但对于更小的模型则效果不明显。
- **逐步推理**[108]：逐步推理是指通过使用思维链 CoT 提示策略，大语言模型可以解决涉及多个推理步骤的复杂任务，如数学问题。CoT 提示利用中间推理步骤来帮助模型得出最终答案，这种能力可能是通过在代码上进行训

练获得的。实验表明，应用大小超过 600 亿的模型时，CoT 提示可以提高模型在算术推理基准任务上的性能。

（5）**评测**：LLMs 评测[109]的一项代表性工作是语言模型的全面评估方法 HELM[110]，这种方法对自然语言处理领域的多种任务场景进行分类，主要考虑 7 种评测指标，包括准确率、校准度、泛化能力、适配能力、鲁棒性、效率和偏见与刻板印象。准确率指的是模型的预测结果和真实结果之间正确比例，判别类问题一般选择 F1 值和 Accuracy 值作为评测指标，而生成任务的指标通常是 BLEU 和 ROUGE。校准度主要评估模型预测的置信度，这有助于提高模型的可靠性，一般的评测指标为期望校准误差（ECE）。泛化能力的评测主要关注模型在训练集之外的测试集上的性能表现，这种评估主要是在零样本或小样本环境下进行的。适配能力是评估 LLMs 的通用任务迁移能力，这种能力主要通过设计提示（prompt）和上下文学习（ICL）来实现的。鲁棒性是评测 LLMs 的抗干扰能力，观察 LLMs 在对抗扰动和非对抗扰动的场景下的性能损失。LLMs 的效率是一个重要评测维度，效率的评测手段通常包括参数个数、FLOPS（运行给定实例模型所需的操作数）、实际推理时间、执行层数等。最后，偏见和刻板印象是生成大模型的重要评估尺度，主要基于表示端评估和生成端评估。表示端的评估主要利用上下文嵌入关联测试（CEAT），基于生成端的评估包括生成内容统计以及生成概率估计。

2.3.2 大语言模型的版式感知方法

大语言模型的版式感知是视觉富文档理解的重要基础。目前，基于纯文本预训练的 LLMs 已经被尝试应用在 VRDs 理解的多种任务上。ICL-D3IE[6]是一种比较经典的方法，该方法提出了一个简单有效的上下文学习框架，通过构建布局感知的上下文范例有效提升了视觉富文档的信息抽取任务。下面简要介绍一下该方法。

图 2.6 展示了模型的细节，此模型包含了两个重要模块，第一个是最近邻文档的选择，利用 Sentence-BERT 表示文档中的文本段，再进行最近邻的比较，找出最相似的多个文档作为候选的上下文范例；第二个部分是多样的上下文范例构建，依次构建硬范例、布局感知的范例和模板范例，并进行迭代更新提示。

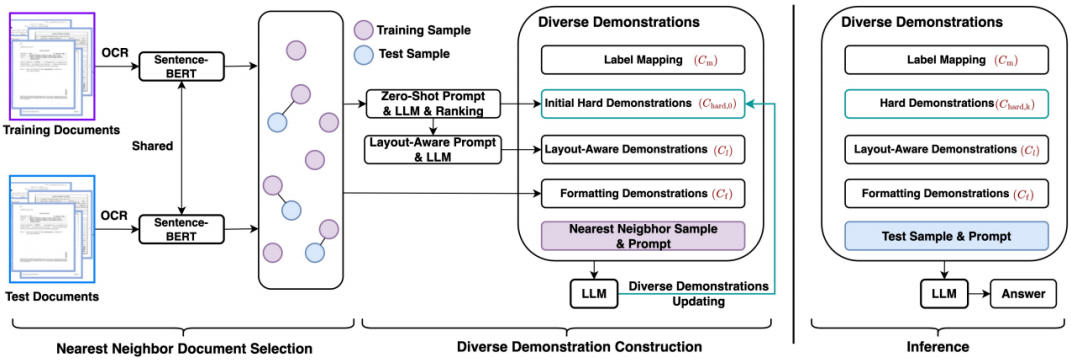


图 2.6　ICL-D3IE 框架图[6]

从整体框架中可以看出，ICL-D3IE 的版式感知能力主要源自多样的上下文范例，作者构建了布局感知的范例，证明了 LLMs 具备版式感知的潜力。实验表明，ICL-D3IE 在多个数据集的小样本场景中都取得了最先进的结果。

然而，目前这个领域研究还很少，并且主要集中在设计上下文范例来提升 VRDs 的性能。针对现在的 LLMs 版式感知方法，本文提出三个缺点和局限性：(1) 缺少统一的版式感知能力的评估方法；(2) 仅仅通过上下文范例进行版式学习；(3) LLMs 版式感知能力难以解释。这些问题还需要进一步的思考和探索。

2.3.3　视觉语言大模型

视觉富文档的理解任务是一个多模态理解任务，这有望通过目前的视觉语言大模型来解决。目前的视觉语言大模型有微软的 LLaVa[111]，智谱 AI 和清华 KEG 发布的 CogVLM[112]，这两个模型都是基于视觉和语言信息融合所提出的。LLaVa 通过对大量文本和图像数据进行训练，使模型具备了强大的语言理解和图像解析能力。在训练过程中，LLaVA 将文本和图像数据转换为向量表示，然后通过计算向量之间的相似度来实现跨模态的匹配和关联。这种匹配和关联使得 LLaVA 能够根据给定的图像生成相应的文本描述，实现图生文。而 CogVLM 在 LLMs (EVA2-CLIP-E) 的基础上，通过添加一个可训练的视觉专家模块来实现视觉信息与语言信息的深度融合。这种方法旨在解决传统视觉语言模型 (VLM) 中的视觉特征与语言模型权重不匹配的问题，从而提升模型在多模态任务上的性能。CogVLM-17B 是目前多模态权威学术榜单上综合成绩第一的模型，在 14 个数据集上取得了第一或者第二名的成绩。图 2.7 比较了不同多模态模型与 CogVLM 在各种任务上的结果。

然而，目前的视觉语言大模型专注在图生文的任务上，虽然已经具备了较好的图文跨模态对齐能力，但是在 VRDs 理解任务中，关键的是文本、排版和视觉信息的跨模态对齐，这部分的研究还值得进一步探索。针对现在的视觉语言大模型与 VRDs 结合的挑战，本文提出两个个问题：（1）视觉和版式的深度融合和对齐；（2）视觉编码器和文本编码器能力的平衡。这些问题值得进一步探讨。

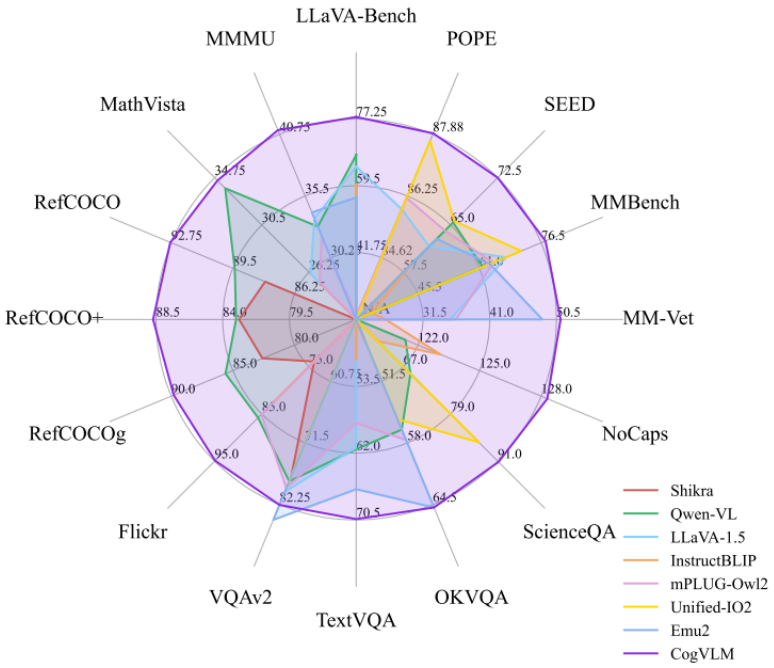


图 2.7 视觉语言大模型在多模态任务上的性能对比[112]

第三章　基于视觉非对称学习的文档实体语义表示方法

3.1　研究动机

视觉富文档（VRDs）的信息抽取（IE）旨在处理各种类型的半结构文档，如发票、收据和表单，这些文档可能是通过扫描或数字化生成的。这一任务引起了学术和工业界的广泛关注[5,113-116]。如图 3.1 所示，VRDs 的 IE 目的是从文档的序列化 OCR（光学字符识别）输出中识别和抽取有意义的语义实体，如公司/个人名称、日期/时间和联系信息。由于单一的文本内容理解难以捕获文档中蕴含的所有语义信息，因此利用多模态信息，包括文本、空间和视觉特征获取更丰富的语义表示是重要的。近年来，大规模预训练的多模态模型（Large Multimodal Models，LMMs）[3-4,9,114-115,117-118]已成为 VRDs 的 IE 中的主要方法，这些模型可以处理多种模态的数据。最先进的 LMMs 将前沿的计算机视觉模型[58,119] 集成

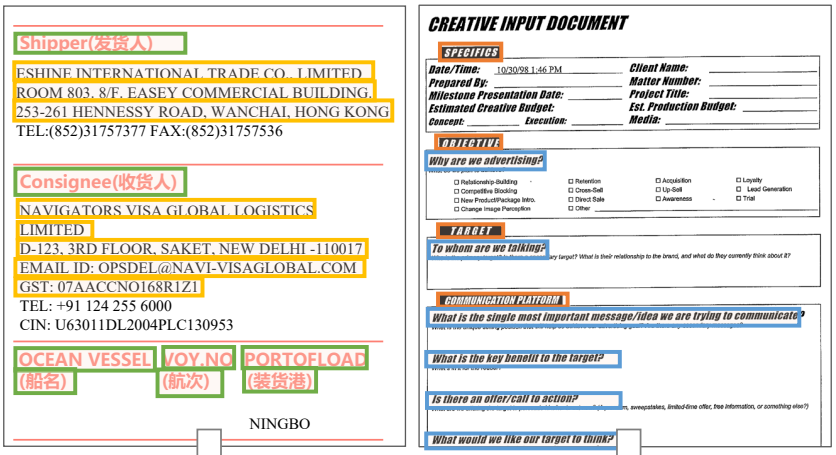


图 3.1　本章工作的动机

到类似于 BERT[73]的架构中，以利用空间和视觉信息以及文本，并学习用于视觉富文档的多模态融合表示。然而，这些表示更偏向于文本和空间模态[120]，当

文档包含更丰富的视觉信息时，这些模型的性能将会受限。这是因为这些模型中的视觉编码器通常与先进的文本编码器相比发挥较次要的作用。

由于相同类型的语义实体通常具有类似的视觉和位置特征，例如相同或相似的字体、背景、颜色，以及边界框的位置和大小，这为识别实体及其类型提供了重要的线索。尽管这些特征十分重要，但现有的用于 VRDs 中信息抽取的 LMMs 往往依赖于一个受限的视觉编码器，无法完全捕获这些细粒度的特征。因此，这项工作专注于通过颜色块将这些视觉先验引入 VRDs 的信息抽取任务中。

先前 LMMs 中的视觉编码器存在两个问题。首先，这些模型只在预训练阶段学习粗粒度的跨模态对齐（例如，文本-图像、单词-单词块和文本-布局对齐[3-4,9]）来增强从视觉通道抽取的特征，但这并不能捕获足够细粒度的视觉特征，并且这些方法也未充分利用视觉先验知识。其次，先前 LMMs 中的视觉编码器的表示能力比最新的光学字符识别（OCR）引擎中的视觉编码器更弱，因为它们不考虑诸如文本段检测和边界框回归等中间任务，这些任务对准确定位和抽取细粒度的视觉特征非常重要。本章的目的是解决这两个问题。

3.2 相关工作

视觉富文档理解：为了实现视觉富文档理解的目标，需要充分利用来自文本、布局和视觉等多种模态的信息来学习语义丰富的预训练语言表示。研究人员提出了各种范式，如基于文本、基于网格、基于图和基于 Transformer 的方法。**基于文本的方法**，例如 XLM-RoBERT[121]、InfoXLM[122]，仅考虑文本模态，依赖于大规模预训练语言模型的表示能力。**基于网格的方法**，例如 Chargrid[89]、BERTGrid[101] 和 ViBERTgrid[91]，使用二维特征图表示文档，然后使用先进的视觉模型进行图像分割和目标检测。**基于 GNN 的方法**[92,123] 将文档中的文本段构建为图结点，并使用图神经网络对文本段之间的关联关系进行建模。**基于 Transformer 的方法**利用最新的多模态预训练模型，通过捕获多个模态的信息来为 VRDs 学习更好的语义表示[124-125]。LayoutLM[5] 基于 BERT[73] 引入了基于二维（2D）相对位置信息，使模型能够感知整个文档中文本段的二维位置。LayoutLMv2[3]、StrucText[116]、StrucTextv2[126] 和 LayoutLMv3[4] 进一步将视觉通道输入集成到统一的多模态 Transformer 框架中，以融合文本、视觉和空间特征。

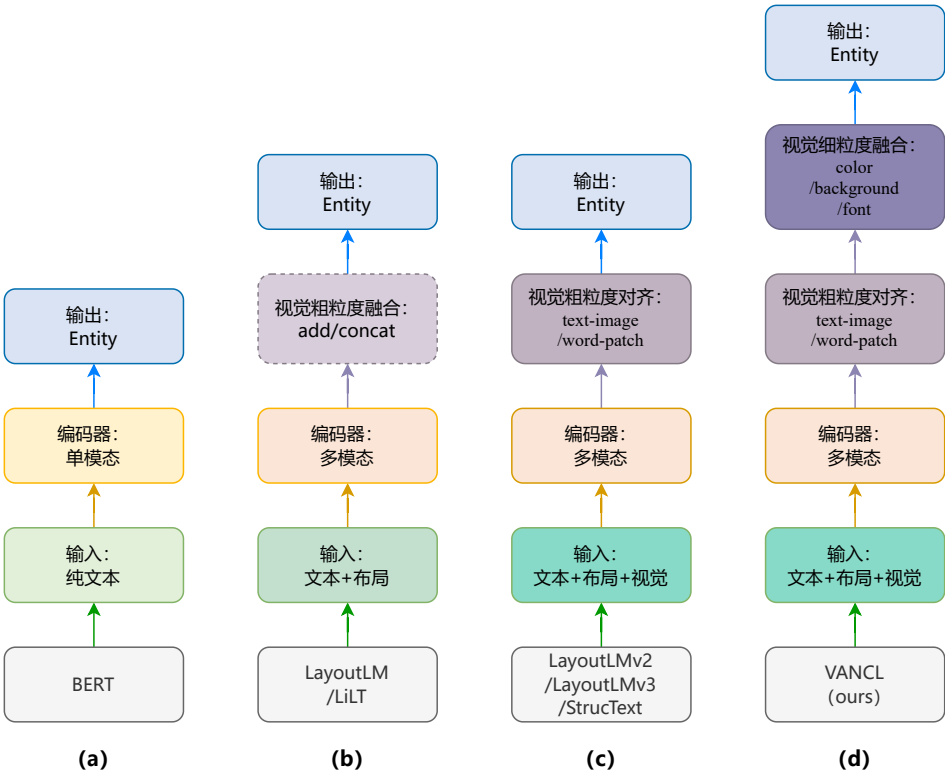


图 3.2 基于 Transformer 架构的多模态文档理解模型的框架对比

因此，这些方法学习到更加健壮的多模态表示，并在视觉富文档理解任务中取得了里程碑的结果。最近，Wang 等人[9] 提出了一个双向注意补充机制（BiACM），以实现内部语言模型之间的跨模态交互。Lee 等人[118] 利用了 OCR 引擎输出的文本段的序列化顺序。基于 Transformer 的方法发展趋势如图 3.2 所示。

一致性学习：近年来，一致性学习领域取得了重大进展，这种学习方式旨在减少不同模型预测之间的差异。一致性学习的核心思想是构建跨视图约束，增强不同模型预测概率分布之间的一致性。先前关于一致性学习的研究集中在对抗性或半监督学习上，这提供了有助于模型从未标记数据中学习的监督信号[127-129]。近年来，有许多工作关注在将一致性机制纳入监督学习中[130-131]。Batra 等人[132] 提出了合作学习，允许多个网络在不同环境中使用不同数据源学习相同的语义概念，这种方法对语义漂移具有抵抗力。Zhang 等人[133] 提出了一种受合作学习启发的深度互相学习策略，使多个网络能够相互学习。

先前的一致性学习研究通常使用经典的知识蒸馏方法[134]，或者使用模态缺陷的生成对抗网络[135] 来实现跨模态学习。Hinton 等人[134] 的方法从一个大规模的、强大的预训练教师网络开始，然后将知识单向传递给一个简单的学生网络，

而 Ren 等人[135] 的方法试图鼓励局部多模态流与完全多模态流一样强大。相比之下，VANCL 探索了跨模态学习和信息传递的潜力，并通过使用视觉线索增强模型的信息抽取能力。这使得模型在遇到特定任务或数据之前就能够获取实体类型的先验知识。在模型结构方面，VANCL 的模型架构更类似于互相学习或合作学习。

3.3 任务定义

语义实体识别（SER）任务视为一个序列标记问题。给定一个由扫描图像 $\mathcal{I}$ 和 OCR 边界框内文本段的列表 $\mathcal{B} = \{b_1, \ldots, b_N\}$ 组成的文档 $\mathcal{D}$，SER 任务被规定为找到一个抽取函数 $F_{\text{IE}}(\mathcal{D} : \langle \mathcal{B}, \mathcal{I} \rangle) \rightarrow \mathcal{E}$，该函数预测了 $\mathcal{D}$ 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” 方案–{*Begin, Inside, Other*} 和一个预定义的标签集。基于预训练的 LMMs 微调一个序列标记模型，并在测试阶段进行维特比解码来预测标记的类型。每个边界框 b_i 包含由 OCR 引擎转录的 M 个标记和定义为 (x_i^1, x_i^2) 的左右水平坐标以及 (y_i^1, y_i^2) 的顶部和底部垂直坐标。

3.4 提出模型

为了解决细粒度视觉特征抽取和无偏多模态表示的问题，受到最近一致性学习的研究工作的启发[127-129,133,136-137]，本章提出了一种新颖的视觉增强训练方法，称为视觉非对称一致性学习（**V**isually-**A**symmetric co**N**sisten**C**y **L**earning，**VANCL**）。通过将颜色先验与基于类别的颜色块作为额外的提示来捕获视觉和布局特征，VANCL 可以增强 LMMs 中无偏多模态表示的学习，即模态均衡。该方法旨在共同考虑文本段（或边界框）检测任务的训练目标和实体类型预测目标，从而将预处理的 OCR 阶段与下游信息抽取阶段联系起来。

VANCL 包括标准学习流程和额外的视觉增强流程，它们的输入在视觉模态上是不对称的，标准流程的输入为原始文档图像，而视觉增强流程中输入的是着色的文档图像。在模型推理时，VANCL 可以将视觉增强流程分离。在训练过程中，VANCL 通过一致性学习来鼓励内部视觉编码器和外部视觉编码器具有同样的表征能力。因此，VANCL 的性能优于现有方法，同时此工作的优势在于：（1）创建着色文档图像所需要的手动工作极少，（2）模型不需要从头开始训练，可以

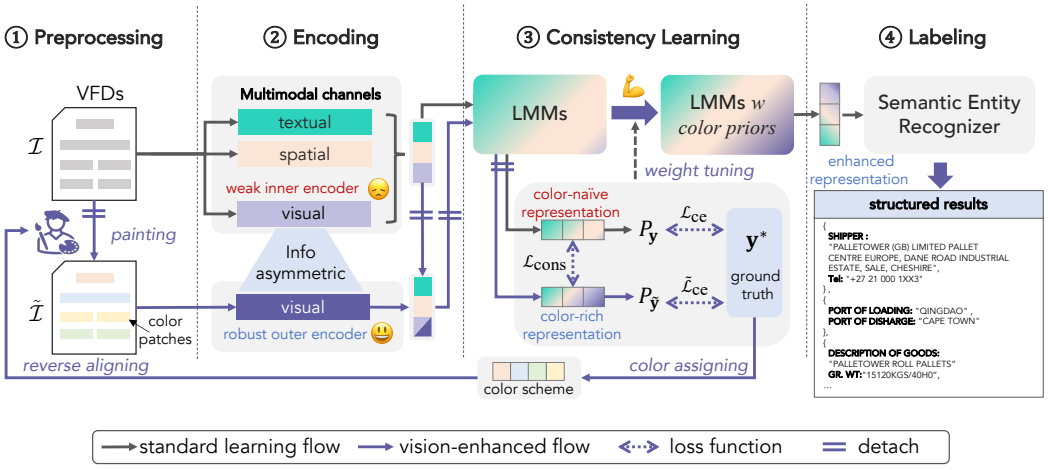


图 3.3　VANCL 的整体框架图

适用于多种主干网络，（3）部署模型的大小不会增加。

视觉非对称一致性学习模型架构如图 3.3 所示。VANCL 模型使用标准学习流程和额外的视觉增强流程，这两个流程同时学习，以将知识从视觉增强流程传输到标准学习流程中。值得注意的是，视觉增强流程的输入图像是含有彩色提示的着色图像，而标准流程的输入图像是原始图像。因此，视觉增强流程的视觉输入和标准流程的视觉输入中包含的信息是不对称的。

3.4.1　整体框架

VANCL 包括两个实体识别网络：一个标准学习（SL）流程用于训练主干 LMMs，另一个额外的视觉增强（VE）流程用于传输先验视觉知识。这两个网络具有相同的结构，并在主干 LMMs 中共享参数权重 Θ。需要注意的是，视觉增强流程具有额外的视觉编码器，因此在训练过程中具有额外的参数权重 Θ_v。这两个流程被形式化为：

$$P_Y = f_{\mathrm{SL}}(X;\Theta), \tag{3.1}$$

$$P_{\tilde{Y}} = f_{\mathrm{VE}}(\tilde{X};\Theta,\Theta_v), \tag{3.2}$$

其中 P_Y 和 $P_{\tilde{Y}}$ 分别是标准学习流程和视觉增强流程输出的预测概率分布，它们是经过 softmax 标准化后的隐含输出（即软标签）。需要注意的是，两个网络的输入是不同的，即 X 和 $\tilde{X}$，前者是原始文档图像，后者是合成的文档图像（带有额外的颜色块）。

训练目标包含两个部分：监督损失 $\mathcal{L}_{\text{sup}}$ 和一致性损失 $\mathcal{L}_{\text{cons}}$。监督损失使用标准的交叉熵损失，如下定义：

$$\begin{aligned}\mathcal{L}_{\text{sup}} = & \frac{1}{|\mathcal{D}^K|} \sum_{\mathbf{x}\in\mathcal{D}^K} \mathcal{L}_{\text{ce}}\Big(\big(P(\mathbf{y}|\langle\mathcal{B}^k,\mathcal{I}^k\rangle;\Theta),\mathbf{y}^*\big) \\ & + \tilde{\mathcal{L}}_{\text{ce}}\big(P(\tilde{\mathbf{y}}|\langle\mathcal{B}^k,\tilde{\mathcal{I}^k}\rangle;\Theta,\Theta_v),\mathbf{y}^*\big)\Big),\end{aligned} \tag{3.3}$$

其中 $\mathcal{L}_{\text{ce}}$ 指的是交叉熵损失函数，$\mathbf{y}^*$ 指的是真实标签。$P(\mathbf{y}|\langle\mathcal{B}^k,\mathcal{I}^k\rangle;\Theta)$ 和 $P(\tilde{\mathbf{y}}|\langle\mathcal{B}^k,\tilde{\mathcal{I}^k}\rangle;\Theta,\Theta_v)$ 分别是标准模型和视觉增强模型的相应预测概率分布。$\mathcal{B}^k,\mathcal{I}^k$ 表示第 k 个文档的边界框位置信息和原始图像。$\tilde{\mathcal{I}^k}$ 是带有颜色块的着色图像。

一致性损失定义了两个预测概率分布之间的相似程度。在训练过程中，由于标准学习流程和视觉增强流程的输入存在不对称信息，因此存在不一致性。具体而言，需要惩罚标准流程和视觉增强流程生成的两个软标签信号（即预测分布）之间的差距。一致性损失计算如下：

$$\begin{aligned}\mathcal{L}_{\text{cons}} = & \frac{1}{|\mathcal{D}^K|} \sum_{\mathbf{x}\in\mathcal{D}^K} Q\big(P(\mathbf{y}|\langle\mathcal{B}^k,\mathcal{I}^k\rangle;\Theta), \\ & P(\tilde{\mathbf{y}}|\langle\mathcal{B}^k,\tilde{\mathcal{I}^k}\rangle;\Theta,\Theta_v)\big),\end{aligned} \tag{3.4}$$

其中 Q 是衡量两个分布之间的差异的距离函数。

最终的视觉非对称一致性学习的训练目标定义为：

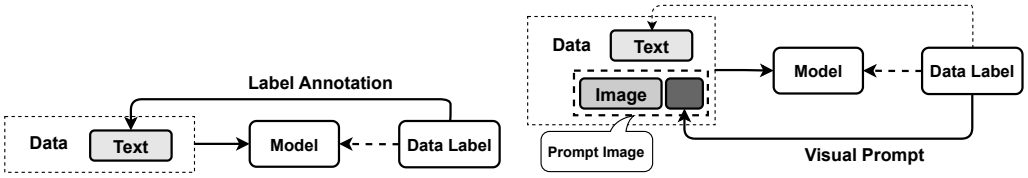
$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{sup}}(\mathbf{y}^*|\Theta,\Theta_v) + \lambda\mathcal{L}_{\text{cons}}(P_{\mathbf{y}},P_{\tilde{\mathbf{y}}}), \tag{3.5}$$

其中 λ 是权衡权重的超参数。上述损失函数考虑了硬标签和软标签之间的一致性，同时减少了模型的过度自信。

3.4.2 基于颜色块的视觉提示

视觉-排版的跨模态对齐能有效地帮助多模态表示的学习，但是边界框级别的细粒度对齐在先前的模型中尚未被充分捕获。因此，探索一种方法来消除文本段（或边界框）检测与实体分类任务之间的差距至关重要。

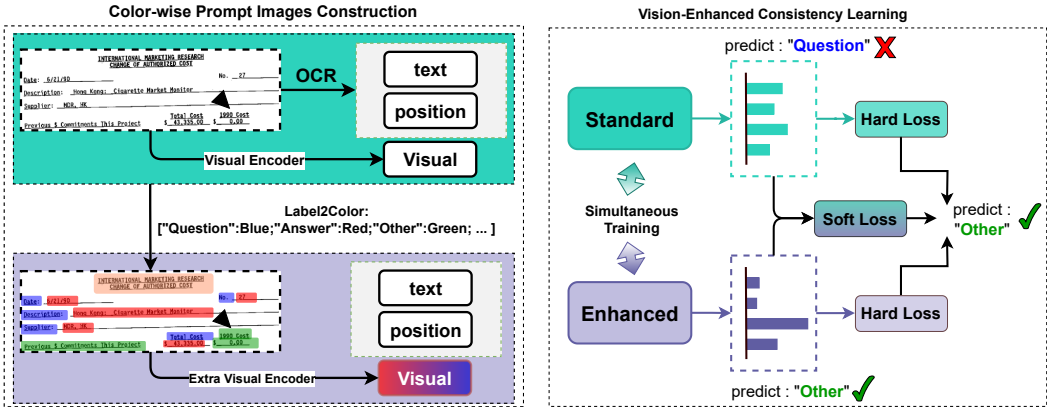
受到 Yang 等人[138]工作的启发，本工作采用一种类似于标签注释（Label



(a) 文本注释的文本提示方式　　(b) 本文的视觉提示方式

图 3.4　文本提示方式与本工作方式对比

Annotation）的视觉提示方式，即在视觉上使用颜色集成标签信息来作为视觉提示，如图 3.4 所示。通过用颜色块表示实体类型信息，可以有效地增强视觉-排版对齐。两者的对比如图 3.5（a）所示。然而，手动创建这些视觉提示将是非常耗时和繁琐的。为了解决这个问题，VANCL 采用了一种简单而巧妙的过程，利用 OCR 边界框坐标自动对原始图像副本中的边界框进行按类别着色。



(a) 基于颜色块的视觉提示模块　　(b) 视觉增强的一致性学习模块

图 3.5　VANCL 模块的具体细节

设 $\mathcal{D}^k$ 表示第 k 个文档，由 $\langle \mathcal{B}^k, \mathcal{I}^k \rangle$ 组成。首先，为每个 VRD 训练样本生成一个图像副本 $\mathcal{I}'^k$。然后，根据边界框 b_i 的坐标 $[(x_i^1, x_i^2), (y_i^1, y_i^2)]$，将图像副本中的边界框着色，以对应实体类型，这样就获得了一个带有颜色块的着色图像 $\tilde{\mathcal{I}}^k$。这个过程可以表示如下：

$$\tilde{\mathcal{I}}^k = \text{Paint}\Big[\text{ROIAlign}\big(\langle \mathcal{B}^k, \mathcal{I}'^k \rangle\big)\Big], \tag{3.6}$$

其中 ROIAlign 获取感兴趣区域（这里是指 $\mathcal{B}^k$）内与边界框对应的细粒度图像区域。$\tilde{\mathcal{I}}^k$ 是保留了原始视觉信息以及带有标签信息的边界框先验的新着色图像。这些着色图像被用来训练额外的视觉增强流程。

3.4.3 基于视觉增强流的一致性学习框架

如图 3.5（b）在直接使用标准流程的提示着色图像训练主干 LMMs 时，避免标签泄露成为这项任务的主要关注点。幸运的是，VANCL 模型中使用的双流架构不仅允许在测试时分离视觉增强流程并丢弃视觉提示的着色图像，还能够使用任意的主干 LMMs 和外部视觉通道中的视觉编码器。这种策略通过一种类似对抗性学习的一致性学习策略，避免了标签泄露，并通过增强原始 LMM 的视觉特征来解决先前模型偏向文本和布局模态的偏见问题。这些都是本工作中新奇和有趣的发现。

3.5 实验

3.5.1 数据集介绍

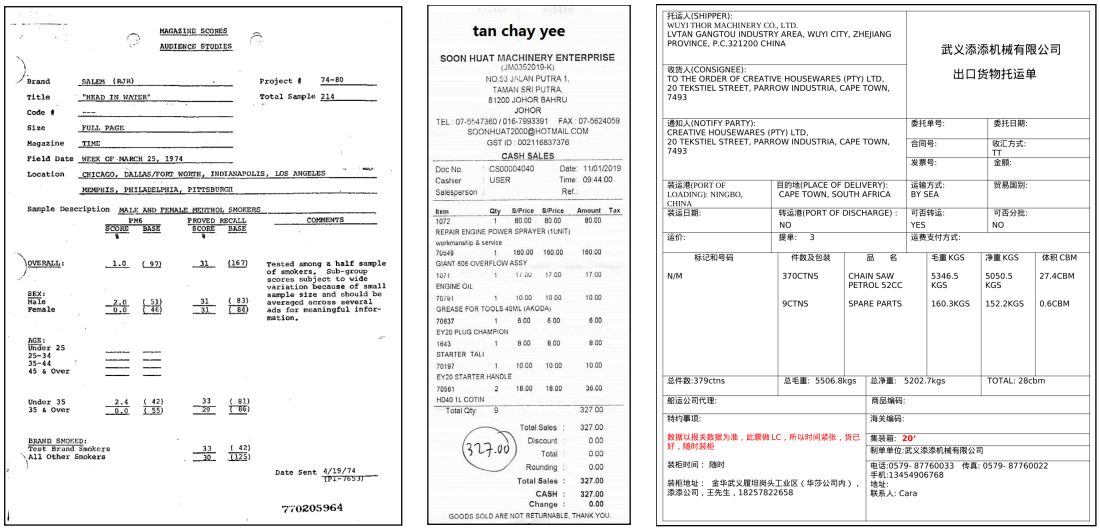
实验数据集：VANCL 在两个开源数据集 FUNSD[139]和 SROIE[95]，还有一个内部收集的 SEABILL 上进行实验。数据集的详细统计信息见表 3.1 。

表 3.1　实验数据集的统计信息

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	21K	33K	50	8K	12K
SROIE	626	34K	155K	347	19K	86k
SEABILL	3,562	250K	1.5M	953	74K	500K

- **FUNSD**：该数据集包含 199 个精心注释的扫描表单。每个语义实体包含一个唯一标识符 id，一个标签 ({问题，答案，标题，或其他})，一个边界框，一个到其他实体的链接列表，以及一个单词列表。具体的样例见图 3.6（a）所示。
- **SROIE**：该数据集包含 626 张用于训练的收据和 347 张用于测试的收据。每张收据包括其扫描图像和 OCR 转录，以文本段列表形式组织，具有边界框位置信息对。每张收据都用四种类型的实体进行标注，即 {公司，日期，地址，总金额}。具体的样例见图 3.6（b）所示。
- **SEABILL**：该数据集是从国际海运场景中提取的一系列复杂文件，包括

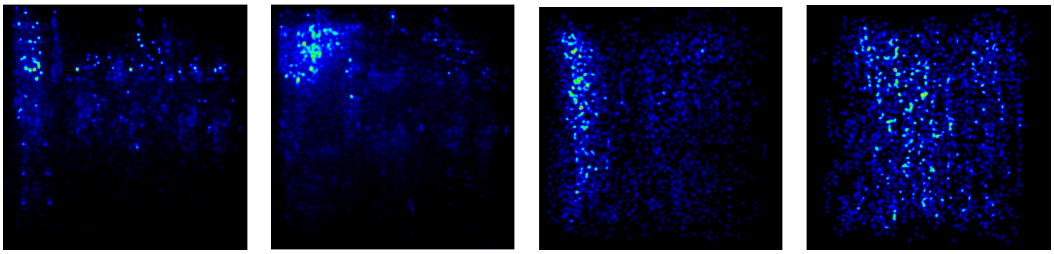
3562 个训练文档和 953 个测试文档。数据由 PDF 图像和基于 PPOCRLabel[①] 标注的文本、位置和类型信息组成，具有粗粒度的三个标签 {问题，答案，其他}，细粒度的 56 个标签 {发货人，发货人-*value*，起运港，起运港-*value*，...}。具体的样例见图 3.6（c）所示。



(a) FUNSD (b) SROIE (c) SEABILL

图 3.6 三个数据集的样例

数据集的实体位置分布：图 3.7 展示了 SEABILL 和 FUNSD 数据集中键和值位置的分布，文档图像被缩放为 1000 × 1000 像素。图中可以看到文档中不同类型实体的分布往往非常明显，例如，键（key）类别的实体分布和值（value）类别的实体分布表现出不一致性。换句话说，这些图表示每个标签类别的先验知识对于视觉富文档信息抽取都至关重要。



(a) SEABILL 的键 (b) SEABILL 的值 (c) FUNSD 的键 (d) FUNSD 的值

图 3.7 SEABILL 和 FUNSD 数据集中键和值位置的分布

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

3.5.2 骨干网络

实验使用 LayoutLM 系列的主干网络进行，这是一系列基于 Transformer 预训练的视觉富文档理解模型。

LayoutLM：基于文本和布局信息预训练的基本 LayoutLM 模型 layoutlm-base-uncased[5]。

LayoutLM *w/ img*：由于基本的 LayoutLM 模型没有利用视觉信息，本实验将 LayoutLM 与 ResNet-101[119] 集成，以实现从视觉通道进行特征抽取的能力①。

LayoutLMv2/LayoutLMv3：这两个改进模型主要扩展了视觉的多模态学习，本实验通过将主干替换为 layoutlmv2-base-uncased[3] 和 layoutlmv3-base[4] 进行比较。

Vancl（本章）：从现有的预训练 LayoutLM 主干初始化 Vancl，并在标准学习和视觉增强流程中共享 LayoutLMs 的参数权重，包括文本编码器（例如 BERT）、内部视觉编码器（例如 ResNet、ResNeXt 和 Linear）、位置编码器、Transformer 层和分类器。对于外部视觉编码器，默认使用 ResNet-101，并测试没有进行预训练 CNN 和 ResNet。

3.5.3 实验环境与细节

Vancl 使用默认参数的 Adam 优化器进行所有版本模型的训练，γ 设置为 $(0.9, 0.99)$，没有预热过程。模型的学习率设置为 5×10^{-5}。对于 FUNSD 和 SROIE 数据集，Dropout 设置为 0.3，以防止模型过拟合，而对于 SEABILL 数据集，Dropout 降低到 0.1。训练批量大小设置为 8，并在 NVIDIA RTX3090Ti GPU 上训练所有模型。本实验将模型训练 20 次迭代以达到收敛，并实现更稳定的性能。

3.5.4 实验结果与分析

（1）主要实验结果

表 3.2 展示了 FUNSD、SORIE 和 SEABILL 数据集上的精确度（P）、召回率（R）和 F1 分数。[*] 表示 Vancl 中使用的骨干模型。粗体表示标准训练方法和

① 注意这与[5] 的做法不同，他们使用 Faster-RCNN[58] 作为视觉编码器。

表 3.2 主要实验结果

Model	#Param	FUNSD			SROIE			SEABILL		
		P% ↑	R% ↑	F1% ↑	P% ↑	R% ↑	F1% ↑	P% ↑	R% ↑	F1% ↑
BERT[73]	110M	54.69	61.71	60.26	90.99	90.99	90.99	66.70	68.77	67.72
RoBERTa[140]	125M	63.49	69.75	66.48	91.07	91.07	91.07	64.35	67.76	66.01
UniLMv2[141]	125M	63.49	69.75	66.48	94.59	94.59	94.59	-	-	-
BROS[142]	139M	80.56	81.88	81.21	94.93	96.03	95.48	-	-	-
DocFormer[117]	149M	77.63	83.69	80.54	-	-	-	-	-	-
LiLT[EN-R][9]	-	87.21	89.65	88.41	-	-	-	84.67	86.02	85.64
LiLT[InfoXLM][9]	-	84.67	87.09	85.86	-	-	-	85.19	87.39	86.29
FormNet[118]	217M	85.21	84.18	84.69	-	-	-	-	-	-
StrucTexT[116]	107M	85.68	80.97	83.09	95.84	98.52	96.88	-	-	-
XYLayoutLM[8]	-	-	-	83.35	-	-	-	90.75	91.59	91.17
LayoutLM[5]	113M	75.97	81.55	78.66	94.38	94.38	94.38	86.93	89.16	88.03
LayoutLM(*w/ img*)	147M	79.68	80.74	80.20	95.53	96.08	95.80	87.50	89.89	88.68
Vancl[LayoutLM(*w/ img*)]	+0M	**80.78**	**81.89**	**81.33**	**96.32**	**96.67**	**96.50**	**89.06**	**90.24**	**89.65**
LayoutLMv2[3]	200M	80.29	**85.19**	82.76	96.25	96.25	96.25	90.90	91.73	91.31
Vancl[LayoutLMv2]	+0M	**82.95**	83.29	**83.12**	**97.45**	**97.58**	**97.51**	**91.61**	**92.17**	**91.89**
LayoutLMv3[4]	133M	-	-	90.29	96.59	96.94	96.77	89.53	91.78	90.64
Vancl[LayoutLMv3]	+0M	**91.76**	**92.95**	**92.35**	**97.09**	**97.30**	**97.20**	**90.64**	**92.07**	**91.35**

Vancl 之间模型性能比较时更好的分数。#Param 指部署时的参数量。Vancl 在精确度、召回率和 F1 分数方面均优于基线模型，包括 LayoutLM（仅考虑文本和空间特征的模型）以及 LayoutLM *w/ img*、LayoutLMv2、LayoutLMv3（还考虑视觉特征的模型）。值得注意的是，LayoutLM *w/ img* 在空间感知多模态 Transformer 层之后合并视觉特征，而 LayoutLMv2 和 LayoutLMv3 在这些层之前合并视觉特征。这表明 Vancl 可以很容易地应用于大多数现有 LMMs 的视觉富文档理解和信息抽取任务，而对网络架构几乎没有重大修改。

正如表 3.2 中所示，在最后一行，Vancl[LayoutLMv3] 在 FUNSD 数据集上实现了最佳性能，超越之前基线模型的 SOTA 结果。值得注意的是，在表 3.2 中可以观察到在不同数据集上的不同程度的提升。Vancl 将 F1 分数显著提高，分别超过了 LayoutLM（*w img*）1.13、0.70 和 0.97 个百分点，超过了 LayoutLMv20.36、1.26 和 0.58 个百分点，超过了 LayoutLMv32.06，0.43 和 0.71 个百分点。这表明 Vancl 成功地从外部视觉增强流程中学习了有用的知识。尽管使用不同的骨干模型和不同的数据集进行的 Vancl 实验结果各不相同，但考虑到 FUNSD 和 SROIE 在数据大小和布局结构密度方面远远小于 SEABILL 数据集，这样的结果是合理的。

Vancl 优于 LayoutLM 系列模型是因为 LayoutLM 仅考虑了粗粒度的视觉-文

本跨模态对齐，例如，单词块-词或图像-文本，它们尚未在边界框级别实现细粒度的跨模态对齐。此外，VANCL 结合了视觉中标签信息的先验知识，因此它学习了更多的视觉特征并增强了潜在的多模态表示。

（2）一致性损失的消融实验

为了对每个模块进行全面评估，本文进行了额外的消融实验，检查使用一致性损失的影响。表 3.3 展示了不同主干网络的一致性损失消融实验，其中-CL 表示未使用一致性损失。表中的实验结果说明，去除一致性损失会导致模型性能在不同程度上的下降。这一发现表明了一致性损失在模型中的重要性。同时，它也表明，没有一致性的驱动力，添加颜色信息会增加学习中的噪音。

表 3.3　不同主干网络的一致性损失消融实验

Model	**FUNSD**	**SROIE**	**SEABILL**
LAYOUTLM(*w/ IMG*)			
+ VANCL	**81.33**	**96.50**	**89.65**
+ VANCL-CL	79.36(↓ 1.97)	95.68(↓ 0.82)	88.19(↓ 1.46)
LAYOUTLMV2			
+ VANCL	**83.12**	**97.51**	**91.89**
+ VANCL-CL	81.52(↓ 1.60)	97.08(↓ 0.43)	89.28(↓ 2.61)
LAYOUTLMV3			
+ VANCL	**92.35**	**97.20**	**91.35**
+ VANCL-CL	90.93(↓ 1.42)	96.98(↓ 0.22)	90.87(↓ 0.48)

表 3.4　不同一致性损失对 F1 分数的影响

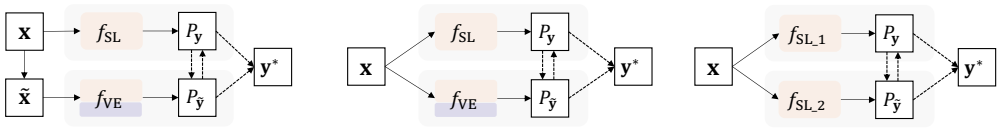
Model	**FUNSD**	**SROIE**	**SEABILL**
LAYOUTLM(*w/ img*)	80.20	95.80	88.68
+VANCL-JS	80.75	**96.50**	89.36
+VANCL-KL	**81.33**	96.32	**89.65**
LAYOUTLMV2	82.76	96.25	91.31
+VANCL-JS	82.21	97.33	**91.89**
+VANCL-KL	**83.12**	**97.51**	91.58
LAYOUTLMV3	90.29	96.77	90.64
+VANCL-JS	**92.35**	97.07	91.16
+VANCL-KL	91.72	**97.20**	**91.35**

（3）不同散度的选择

衡量两个分布之间差异的方法有多种，不同的衡量标准可能导致不同的一致性损失，这也会影响模型在视觉富文档抽取任务上的最终结果。本节验证了两种类型的一致性损失，即 Jensen-Shannon (JS) 散度[143] 和 Kullback-Leibler (KL) 散度[144]，对实验结果的影响。尽管 VANCL 模型在大多数情况下优于基线模型，无论是不同的主干网络还是数据集，但仍值得注意的是，一致性损失的选择取决于数据集的特征，如键值类型和版式布局结构，以及由于模型复杂度和数据大小导致的过拟合是否发生。如表 3.4 中的结果所示，针对不同数据集和主干网络使用不同的一致性损失以达到最佳结果。例如，当在 LayoutLMv2 中使用 JS 散度时，VANCL 在 SEABILL 数据集上达到了最佳结果，但在 FUNSD 和 SROIE 数据集上只能达到次优的结果。

（4）内/外部一致性的比较

为了检验 VANCL 一致性学习的有效性，本文将 VANCL 与 R-Drop[136] 和 Mutual learning[133] 进行比较。R-Drop 是一种强大且广泛使用的正则化方法，考虑了多个模型内部一致性信号。图 3.8 展示了这三种方法的不同结构，‘→’表示前向操作，‘- ->’表示监督损失。表 3.5 给出了将 LayoutLM 模型与 R-Drop 和 Mutual learning 结合的结果。与 R-Drop 和 Mutual learning 相比，可以发现模型在 FUNSD、SROIE 和 SEABILL 数据集上使用 VANCL 可以显著的提升性能。实验结果表明，通过一致性学习，视觉先验有效的提升了 VANCL 的视觉富文档理解能力。



(a) VANCL　　(b) Mutual learning　　(c) R-drop

图 3.8　三种不同的一致性学习框架对比

（5）视觉编码器共享参数的影响

为了探究视觉编码器权重共享的效果，本文通过控制模型权重共享的变量，以 LAYOUTLM（*w/ img*）为主干网络，在三个数据集上进行实验。如表 3.6 所示，其中“Visual”指的是视觉编码器的参数是否共享，表中结果可以看到，在大多数情况下，共享视觉编码器的参数比不共享权重更好，而且共享参数的方式还大

表 3.5　内部和外部一致性的 F1 分数比较

Model	FUNSD	SROIE	SEABILL
LayoutLM (*w/ img*)	80.20	95.80	88.68
+Mutual-Learn	80.44	95.46	88.99
+R-Drop	80.38	96.02	89.13
+VanCL	**81.33**	**96.50**	**89.65**

大减少了模型的总参数数量（见表 3.2）。此外，本文还比较了在是否共享权重的情况下使用颜色块的着色图像是否会对结果产生影响。即使在模型权重不共享的情况下，颜色视觉线索仍然显著提高了模型的性能。

表 3.6　视觉编码器共享网络参数的实验结果

Visual	Prompt	FUNSD	SROIE	SEABILL
¬ *shared*	✖	77.47	95.13	88.63
	✔	80.03	95.84	89.13
shared	✖	80.38	96.02	89.13
	✔	**81.33**	**96.50**	**89.65**

（6）多模态表示可视化

为了更直观地展示 VanCL 学习的视觉编码在视觉通道中的变化，本文通过 t-SNE 可视化了 VanCL 训练前后的编码分布。图 3.9 展示了两个流程中的视觉信

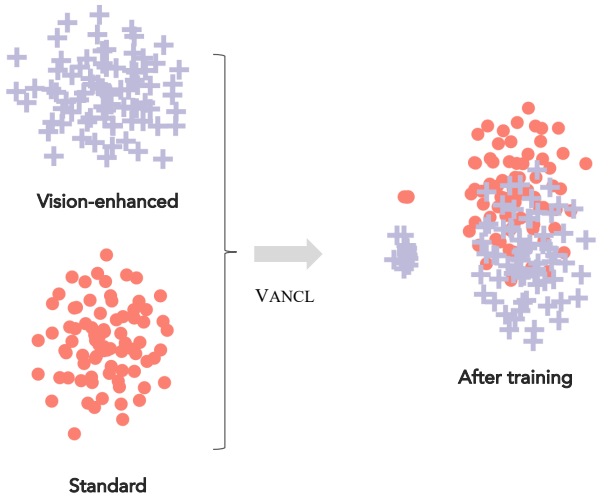


图 3.9　标准流和增强流的视觉编码 t-SNE 可视化

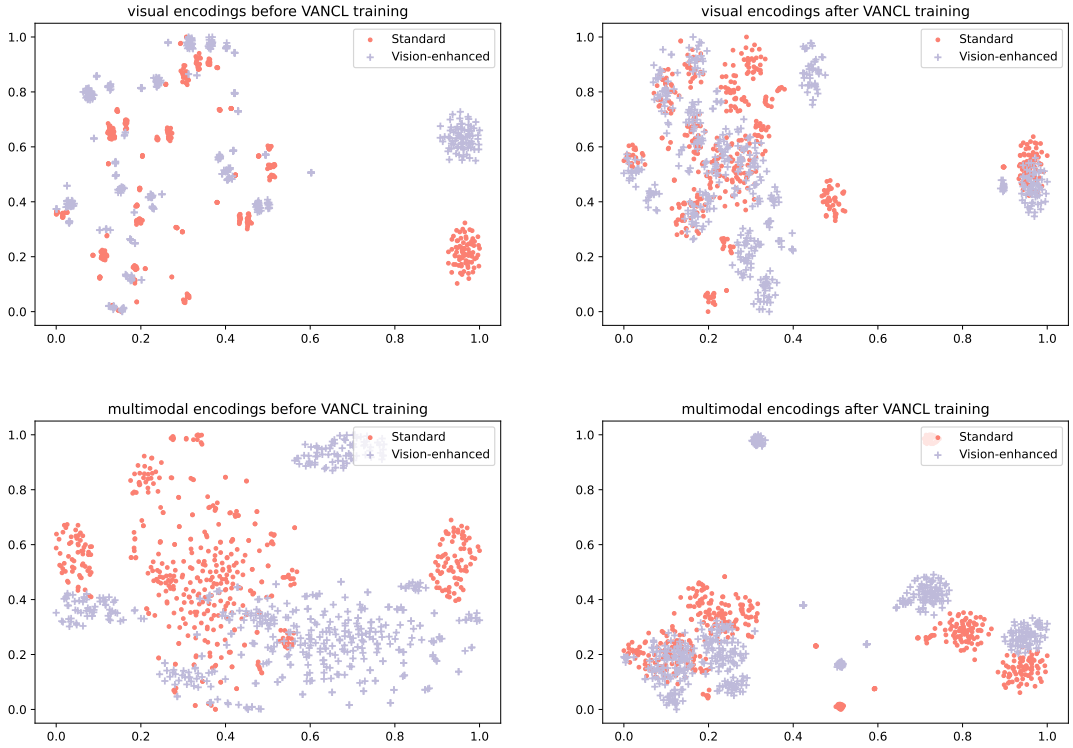


图 3.10 标准流和增强流输出的隐藏状态的完整 t-SNE 可视化结果

息的部分结果。图中表明，VANCL 通过调整标准流程和增强流程中的信息来间接地将标签先验分布传递给标准流，从而提升了后续的推理过程。完整版的 t-SNE 可视化结果如图 3.10 所示。左上和右上图分别展示了 VANCL 训练前后两个流程的视觉编码的比较。训练后的结果展示了每个分类簇更加分离，两个流的对应信息更加对齐。左下和右下图分别展示了 VANCL 训练前后两个流程的多模态输出的比较。总之，VANCL 通过视觉增强流有效地将标签信息传递给了标准流。

（7）色块颜色对结果的影响

为了验证不同背景颜色对模型的影响，本文在 FUNSD 数据集上进行了颜色的消融实验。除了标准的颜色方案（红色、蓝色、黄色和绿色）外，还进行了三个额外的实验。第一，随机交换标准颜色方案中使用的颜色。第二，为每个标签类别选择相同颜色系统的不同色调或强度。第三，仅使用颜色绘制边界框线。如表 3.7 所示，**SUPP.** 表示测试集中每种实体类型的数量。表中的结果展示了 VANCL 遵循的三条规则：1）相同颜色系统的不同色调或强度对结果影响不大，但性能受限；2）交换当前颜色方案中的不同颜色对结果影响不大；3）VANCL 在使用对比度强的颜色时有效且对灰度敏感。这与认知科学中的现有研究结果一

表 3.7　不同颜色方案在 FUNSD 数据集上的结果

	QUESTION		ANSWER		HEADER		OTHER		MICRO-AVG
	color	F1	color	F1	color	F1	color	F1	
1	#FF0000	83.88	#0000FF	**85.51**	#00FF00	56.31	#FFA500	77.36	**81.33**
2	#0000FF	**84.42**	#FF0000	85.49	#00FF00	57.92	#FFA500	75.83	81.28
3	#FFA500	84.01	#00FF00	85.22	#FF0000	54.08	#0000FF	77.38	81.27
4	#CCCCCC	83.67	#999999	84.78	#333333	57.18	#000000	**77.47**	81.16
5	#FF0000	83.52	#FF6699	84.88	#FF3366	56.02	#FF0099	76.23	80.75
6	#0000FF	83.98	#FF0000	84.59	#0099FF	56.40	#0066CC	76.42	80.76
7	#FF0000	83.43	#0000FF	83.62	#FFA500	55.73	#00FF00	76.30	80.23
8	#FFFFFF	83.45	#FFFFFF	84.21	#FFFFFF	**61.40**	#FFFFFF	75.58	80.38
SUPP.	2,542		3,294		374		2,356		8,566

致[145-146]。

（8）低资源场景的比较

为了验证 VANCL 在低资源场景中的有效性，本文选择了不同规模的数据训练模型，观察 VANCL 是否会提高 SER 任务的性能。本实验选择 LayoutLM 和 LayoutLMv2 作为测试主干网络，并通过改变训练数据的大小，比较 VANCL 的结果与相应基线的结果。具体地，从 SEABILL 数据集中随机选择训练数据的百分比 p，其中 p 可以取自集合 $\{5\%, 12.5\%, 25\%, 50\%\}$。

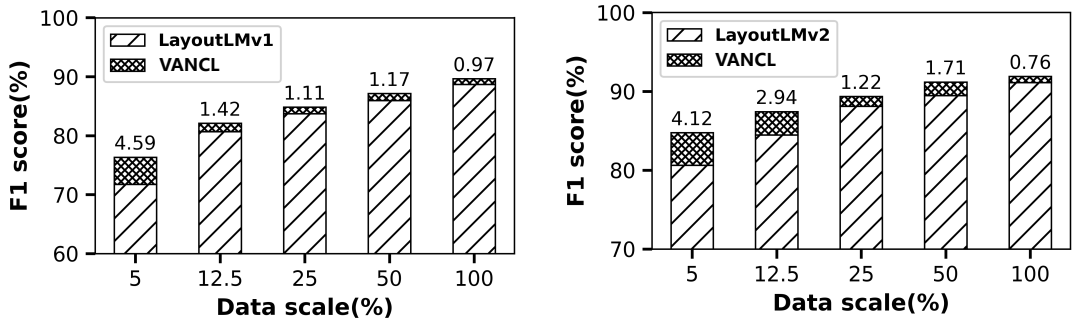
表 3.8　不同规模 SEABILL 训练子集的结果

Model	Ratio (F1)				
	5%	12.5%	25%	50%	100%
LAYOUTLM(*w/ img*)	71.74	80.67	83.72	85.97	88.68
+VANCL	**76.33**	**82.09**	**84.83**	**87.14**	**89.65**
LAYOUTLMV2	80.63	84.48	88.12	89.47	91.31
+VANCL	**84.75**	**87.42**	**89.34**	**91.18**	**91.89**

实验结果如图 3.11 所展示，结合表 3.8 中的结果，可以得到两个结论：（1）VANCL 在不同规模的训练数据上始终优于 LayoutLM 基线。（2）在训练数据集越少的场景下，VANCL 相对基线的性能提升会更大，这表明 VANCL 可以在这种低资源的场景下显著的提升模型训练的效果，增强模型的鲁棒性。

（9）视觉编码器颜色感知的重要性测试

为了验证 VANCL 的改进是否仅仅是因为预训练的视觉编码器包含了真实世



(a) LayoutLM(*w/ img*)　　(b) LayoutLMv2

图 3.11　不同规模 SEABILL 训练子集的柱状图

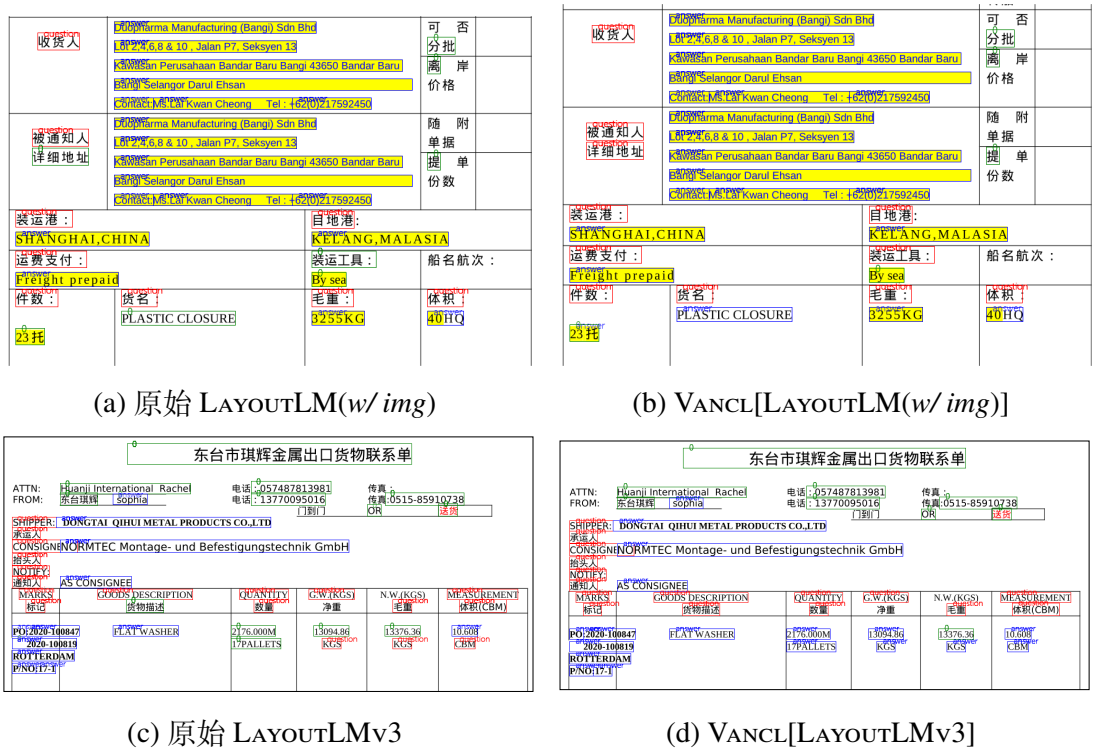
界的颜色先验，本文使用了未经预训练的视觉编码器，并在引入模型时初始化其参数。这样一来，视觉编码器就需要从零开始学习抽取视觉特征的能力。考虑到更深层的网络通常在短时间内更难训练，所以选择了一个较小的 ResNet-18 和一个极简单的两层 CNN[147] 作为实验的外部视觉编码器。如表 3.9 ，其中 Pre. 表示预训练。Enc. 和 #L 分别表示外部视觉编码器和内部层数。从表中可以看到，即使是简单的视觉编码器网络，如 CNN-2 和 ResNet-18，也能够超过标准基线模型的结果，这表明原始的主干模型通过视觉非对称一致性学习（VAncl）学习到更好的视觉特征。

表 3.9　使用未经预训练的外部视觉编码器对 LayoutLM（*w/img*）的影响

Model	Enc.	#L	Pre.	FUNSD	SEABILL
Stand.	-	-	-	78.66	88.03
Stand.	ResNet	101	✔	80.20	88.68
Stand.	ResNet	18	✖	79.51	88.50
stand.	CNN	2	✖	80.03	88.42
Vancl	ResNet	101	✔	**81.33**	**89.65**
Vancl	ResNet	18	✖	77.62	88.64
Vancl	CNN	2	✖	80.60	88.97

（10）案例可视化展示

图 3.12（a）和（b）和图 3.12（c）和（d）分别可视化了 SEABILL 数据集上使用普通 LayoutLM(*w/ img*) 和 LayoutLMv3 模型以及 Vancl[LayoutLM(*w/ img*)] 和 Vancl[LayoutLMv3] 模型的对应预测结果。在图 3.12（a）和（b）中，



(a) 原始 LayoutLM(*w/ img*)　　(b) VanCL[LayoutLM(*w/ img*)]

(c) 原始 LayoutLMv3　　(d) VanCL[LayoutLMv3]

图 3.12　SEABILL 数据集上使用标准微调流程对比 VanCL 流程的预测结果案例研究

LayoutLM(*w/ img*) 对“PLASTIC CLOSURE”、“装运工具 (Loading way)”，“详细地址 (Detailed address)”，“23 托 (23 GP)”等标签预测错误为“*Other*”，而 VanCL[LayoutLM(*w/ img*)] 可以做出与人工标注的真实情况相匹配的正确预测。但也存在一些模糊的情况，即使是 VanCL[LayoutLM(*w/ img*)] 也会给出错误的预测。例如，LayoutLM(*w/ img*) 和 VanCL[LayoutLM(*w/ img*)] 都将“By sea (23 GP)”预测为“*Other*”，但真实标签是“*Answer*”。

图 3.12（c）和（d）显示，LayoutLMv3 比 LayoutLM(*w/ img*) 给出了更准确的预测。然而，LayoutLMv3 也对“货物描述 (Goods description)”，“2176.000M”，“17PALLETS”，“13094.86”等实体预测错误的标签为“*Other*”，尽管这些实体在训练数据中被注释为“*Question*”或“*Answer*”。对于“KGS”，“CBM”，LayoutLMv3 也给出了错误的标签“*Question*”，而 VanCL[LayoutLMv3] 给出了正确的标签。

3.6 本章小结

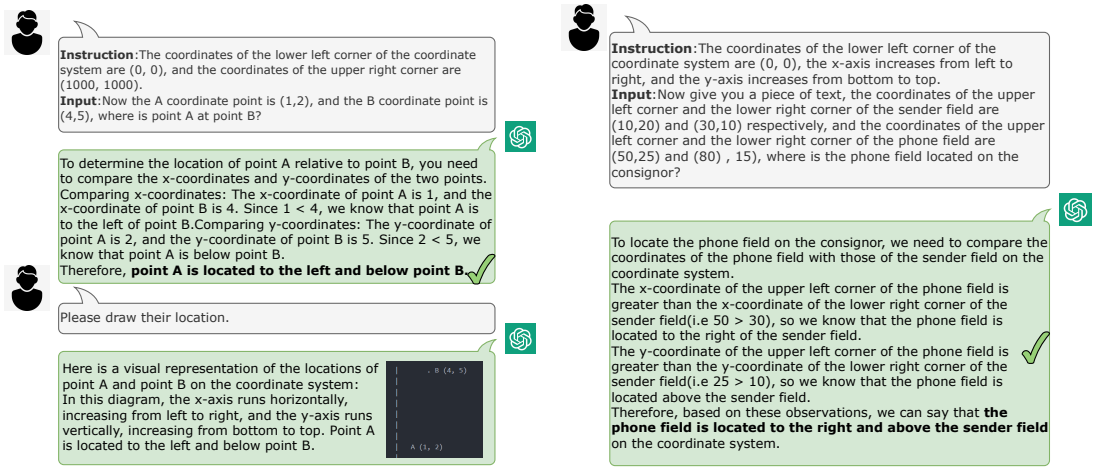
本章提出了一种新颖的视觉-排版跨模态一致性学习方法 VANCL，使用视觉非对称作为输入，这有效地将视觉先验知识融入多模态表示学习中，解决了目前多模态视觉富文档理解模型难以捕捉细粒度视觉特征的问题。基于视觉-排版先验的一致性学习方法解决了 LMMs 中存在的多模态信息有偏的问题，实现了多模态均衡。实验展示了 VANCL 的多模态表示概率分布相对传统的 LMMs 更优。本章将 VANCL 被应用于视觉富文档信息提取的任务，并使用不同的 LMMs 骨干网络来证明了此模型的有效性。实验结果表明，VANCL 与使用的骨干网络无关，并取得了目前最好的结果。

第四章　基于布局提示的大模型文档实体关系理解方法

4.1　研究动机

传统的多模态文档理解模型已经可以解决各种复杂的视觉富文档理解任务，并逐渐向通用的文档智能模型发展。然而，通用的文档智能要求使用统一的模型解决各种文档理解任务，如文档分类，信息抽取，文档问答等，这给传统文档智能模型的多任务适配带来了挑战。同时，为了提高文档智能模型的通用价值，理解自然语言并与人类进行无差别对话是模型设计的诉求。

近期，诸如 ChatGPT①、LLAMA②、CogVLM③ 等大语言模型（LLMs）在大多数自然语言理解任务上取得了显著的成果，通过构建简单的提示模板可以方便地统一整个下游理解任务，并且对话大语言模型天生具有强大的人机交互能力，可以广泛应用到各种实际场景。因此，大语言模型对话是一种更通用、更先进的智能设计理念。此外，随着训练数据和模型参数的呈指数增长，许多研究[108,148]已经证明，当参数量达到一定水平时，LLMs 会展现出之前几乎不存在的能力，这被称为大语言模型的能力涌现。在给模型一些演示示例后，这些能力



(a) 基本位置理解　　(b) 实体位置关系理解

图 4.1　ChatGPT 空间位置感知示例

① https://openai.com/blog/chatgpt
② https://github.com/facebookresearch/llama
③ https://github.com/THUDM/CogVLM

会变得更加有效，这被称为上下文语境学习（In-Context Learning，ICL）[149]。

然而，大多数先前的研究都集中在 LLMs 的文本语义理解能力上[107-108,150]，LLMs 涌现的各种能力，比如文本-布局跨模态理解，是否 LLMs 具有空间感知的能力，尚未深入研究。虽然 LLMs 是基于海量纯文本语料训练的，但图 4.1 却给出了一个有趣的真实例子，即 ChatGPT 不仅能够理解文档页面中的二维位置关系，还具有一定的空间布局生成能力。如图中所示，其生成的位置基本上是符合逻辑的，并且 ChatGPT 在回答的过程中给出了详细的论证步骤，而不是直接给出答案，这表明 ChatGPT 具有强悍的布局感知潜力。然而，由于缺乏相关研究，LLMs 的版式感知和生成能力仍然不确定。

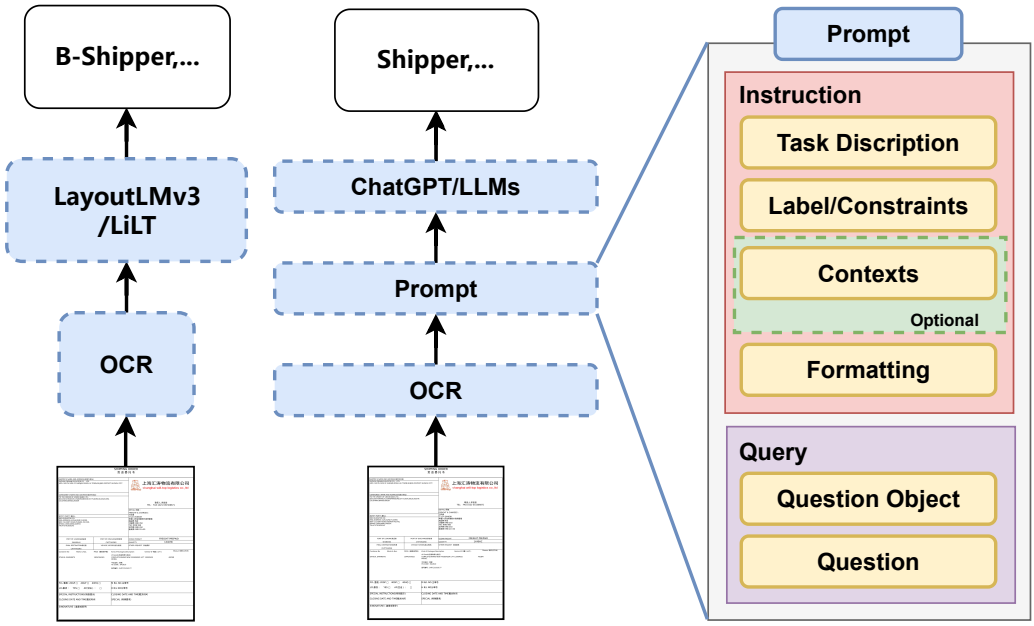
由上述章节可以得知，版式布局是视觉富文档键值关系理解的重要特征，而理解版式布局的前提是理解二维位置关系。因此，本章着重探究 LLMs 的空间位置理解能力，设计了四种基于二维位置理解的评估方案和数据集，从位置关系、位置预测、关系理解和布局生成四个角度客观的评测 LLMs 的空间位置理解水平。在这个基础上，本章提出一种基于布局感知提示的 LLMs 文档实体关系理解方法，实验表明，所提出的方法可以有效的填补了 LLMs 在版式布局理解上的偏差。

4.2 研究方法

（1）框架对比

近年来，多模态文档理解模型在视觉富文档的信息抽取任务上取得了令人印象深刻的结果，这种方法主要依赖“预训练-微调（fine-tuning）”的方式进行领域任务的理解。与这种方式不同，大语言模型的语义理解方式主要是“预训练-提示（prompt）”，这种方式不依赖大量的标注数据，只需要设计足够好的提示就可以诱导模型输出预想的结果。图 4.2 展示了两种方式在文档理解任务上的区别。

先前的预训练文档理解模型，如 LayoutLMv2，LayoutLMv3，需要在全量的训练数据集上进行微调，然后再进行推理过程。而 LLMs 省去了繁琐的微调过程，直接利用模板化的提示作为 LLMs（ChatGPT、GPT4 等）的输入，最后得到预期的结果。在这个过程中，提示的构建起到了决定性的作用，通常的提示（prompt）包含指令（instruction）和查询（query），指令是 LLMs 的前缀（prefix）



(a) 多模态模型　　(b) 大语言模型

图 4.2　现有多模态模型 LMMs 与大语言模型 LLMs 在文档理解任务上的区别

提示，一般由任务描述、标签映射、约束、可选的上下文范例和格式化输出范例组成，而查询是任务所关心的问题，一般是由问题对象（Question Object）和具体问题构成。通过这种方式请求 LLMs 可以获得预期的结构化输出，然后可以进行进一步的处理以供下游任务使用。

（2）提示构建

本小节介绍 LLMs 如何构建合理的提示，包括指令的构建和查询提示的构建两部分。LLMs 提示的基本框架如图 4.2（b）所示。

任务描述（Task Description）：每个任务都需要一个特定的任务描述 I_t，以便 LLMs 专注于特定的任务目标并调用其已知的知识。为了构建更详细的任务描述，描述需要被精心设计，并使用合理的自然语言进行增强，这使得 LLMs 对任务的对象有更多的了解。以语义实体识别（SER）任务为例，一个合理的任务定义描述为：

“你的任务是语义实体识别，对于给定的文本段和边界框的位置，充分理解文本语义和位置关系，然后从标签集合中选择一个最相关的类别作为该语义实体的类别预测”。

这个描述会让 LLMs 扮演执行语义实体识别任务的角色，并调用与其相关的知识解决上下文的问题，这是一种方便且有效的任务迁移方式。

标签映射（Label Mapping）：文档信息抽取任务需要给文档的每一个语义实体字段预测一个类别。因此，在指令中需要指定 LLMs 预测的标签空间（Label Space）。标签映射 I_{l} 的目标是将语义相似的词标签转换为同一个标签，使得 LLMs 能够有效地作为一个分类模型发挥作用。为了实现这一目标，需要从提供的数据集中收集原始标签的文本描述，例如“发货人”代表“发货人和发货人的内容（“-value”后缀表示）”，然后将原始标签（Y'）及其对应的描述（Y）包含在上下文中，以提示 LLMs 解决测试样例。用于提示的标签映射用公式表述为：

$$I_{\text{label}} = \text{CONCAT}\left(Y', Y\right) \tag{4.1}$$

上下文范例（Context Demonstrations）：随着上下文学习[149,151]的发展，可以在进行零样本推理的时候添加这部分指令来作为任务的提示范例。以 VRDs 信息抽取为例，上下文范例通常是一些具有显著特征的代表性样本，例如关键信息排列的固有模式，键值位置关系等。VRDs 的上下文演示通常包含三个部分：第一，一个 OCR 转录，包括文本段 T_{c} 和相应边界框的位置坐标 B_{c}；第二，一个问题 Q_{c}，比如“这些文本段在文档中属于哪个类别?”；第三，问题的正确答案 Y_{c}。此处注意，上下文范例并不是提示中必须包含的，如果包含上下文范例，那么此时 LLMs 进行的是小样本/少样本推理，如果不包含，则是零样本推理。一个基于 VRDs 信息抽取的上下文范例可以被形式化为：

$$I_{ctxt} = \text{CONCAT}\left(T_{\mathrm{c}}, B_{\mathrm{c}}, question_{\mathrm{c}}, Y_{\mathrm{c}}\right) \tag{4.2}$$

其中 Y_{c} 是标签映射集合 I_{l} 中的一个类别。

约束（Constraints）：在进行 LLMs 的推理过程中，不可避免的会因为模型能力太过于广泛而导致无意义的结果。因此，有必要在指令中加入了适当的约束条件 I_{cons}，让模型舍弃掉一些不符合任务要求或者不合理的回答。例如在文档字段位置生成任务中，应该约束模型生成的位置不能与文档中其他字段的位置重叠，因为这是毫无意义的。

格式化输出（Formatting）：根据[6]的建议，LLMs 的输出格式需要保持一致和标准化，这样有助于后续的数据处理与评估，因此有必要给出预期的格式化输出形式 I_{f}。以 VRDs 信息抽取任务为例，LLMs 的输出应该要包含文本段的

ID 索引，文本段内容、边界框位置和最终预测的类别，通常保存成方便处理的 JSON 文件格式。一个格式化输出的例子如下所示：formatting：{idx:xxx, text:xxx, position:[xxx], label:xxx}。

查询提示（Query）：查询提示 Q 的设计与上下文范例相似，唯一不同的是，查询提示中包含的是本次请求的问题对象和具体问题，标签是需要 LLMs 预测的。其被形式化为

$$Q = \text{CONCAT}\left(t_{\text{q}}, b_{\text{q}}, question\right) \tag{4.3}$$

推理（Inference）：在获得了完整的指令和查询提示之后，LLMs 能够根据序列化的提示预测具体的结果，以 VRDs 信息抽取任务为例，预测整个文档所有文本段的类别，这个过程可以表述为：

$$P(Y \mid I, Q) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{P}_{llm}\left(y_i \mid I_{\text{t}}, I_{\text{l}}, I_{\text{cons}}, I_{\text{ctxt}}, I_{\text{f}}, Q\right) \tag{4.4}$$

其中 N 的含义是整个文档中所有语义实体的数量。

4.3 空间位置理解评估

为了客观的评估 LLMs 的二维空间位置理解能力，本小节从相对位置理解、语义位置预测、语义位置关系理解和语义布局生成这四个角度分别提出了四种评估方案，包括基本空间感知（BSP）、页面位置预测（PLP）、文档信息抽取（DIE）和文档布局生成（DLG），其中 BSP 和 PLP 侧重评估 LLMs 的局部和整体的位置理解能力，而 DIE 和 DLG 则侧重评估 LLMs 局部和整体的实体语义关系理解能力。下面分别介绍这四种评估方案。

4.3.1 基本空间感知

任务定义：基本空间感知（Basic Spatial Perception，BSP）要求 LLMs 能够理解坐标的含义，以及坐标之间的相对位置关系。在视觉富文档 VRDs 的理解任务中，位置关系的理解是语义识别和布局理解的前提。换句话说，BSP 就是评估 LLMs 的“方向感”，检查模型是否真正理解上下左右的空间语义，而不仅仅是文本层面的理解。

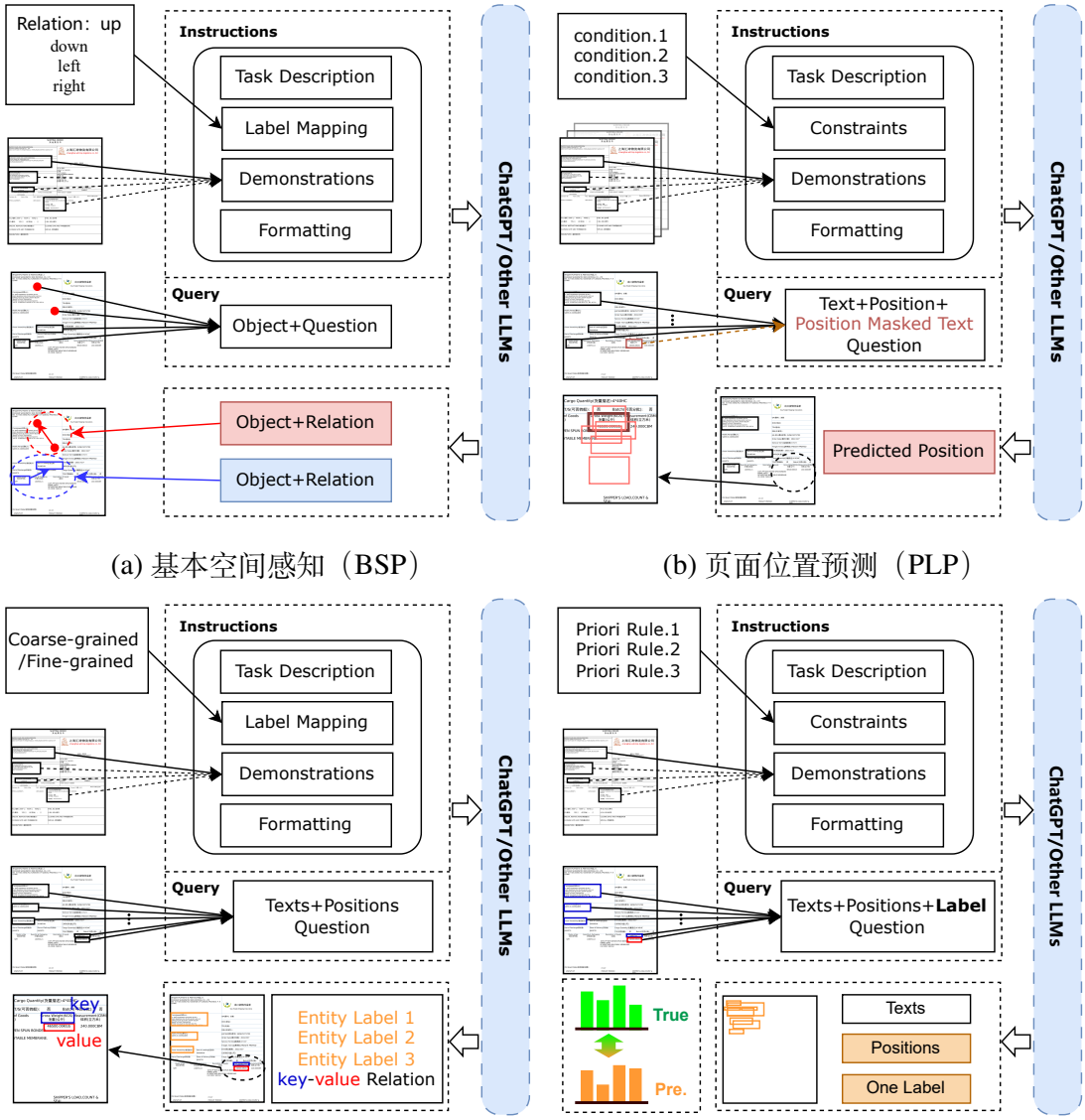
Algorithm 1 计算边界框（bbox）之间的位置关系

输入: 第一个边界框 $\mathcal{B}_1$, 第二个边界框 $\mathcal{B}_2$, 位置关系列表 $\mathcal{S}$
输出: 位置关系 $\mathcal{R}$

$\mathcal{R} \leftarrow None$
if bbox2 被 bbox1 的左、右、上边界线包含 **then**
    $\mathcal{R} \leftarrow above$
else
    if bbox2 被 bbox1 的左、上、下边界线包含 **then**
        $\mathcal{R} \leftarrow left$
    else
        if bbox2 被 bbox1 的左、右、下边界线包含 **then**
            $\mathcal{R} \leftarrow below$
        else
            if bbox2 被 bbox1 的右、上、下边界线包含 **then**
                $\mathcal{R} \leftarrow right$
            else
                计算 bbox1 和 bbox2 中心点的相对位置 $\mathcal{R}_c$
                $\mathcal{R} \leftarrow \mathcal{R}_c$
            end if
        end if
    end if
end if
return 位置关系 $\mathcal{R}$

BSP 包含两个方面的评估任务，即点与点（point-to-point）和边界框与边界框（bbox-to-bbox）之间的相对位置预测。注意，本文所描述的边界框位置都是文本框左上角和右下角坐标唯一确定的。BSP 任务被认为是一个分类任务。对于点来说，查询提示需要包含两个点的位置坐标，LLMs 需要利用自身的二维位置理解能力判断这两个坐标点的相对关系，如左上、左下、右上、右下等，在这个过程中，LLMs 被引导去比较点的横坐标和纵坐标的大小。与点的位置关系比较不同，bbox-to-bbox 的比较更为复杂，由于文本边界框的长宽通常不一致，合理判断两个边界框的位置关系需要 LLMs 对空间方向有一定的感知。因此，BSP 评估任务设计了一个基于 bbox-to-bbox 的位置判断算法，通过自然语言的描述作为 LLMs 的输入，评估模型是否能依据算法规则实现 bbox 级别的位置关系预测。算法的流程如算法 1 所示。

任务流程：图 4.3（a）展示了 BSP 任务的主要流程。任务描述会对 BSP 相对位置判断的规则用自然语言描述，并注明将该任务视作为一个多分类的任务。标签映射包含了可能存在的多种位置关系，以供 LLMs 选择。此处注意，为了保证评估结果的客观性，所有的评估任务都没有使用上下文范例进行提示。格式



(a) 基本空间感知（BSP）　(b) 页面位置预测（PLP）

(c) 文档信息抽取（DIE）　(d) 文档布局生成（DLG）

图 4.3　LLMs 空间位置理解评估方案流程

化输出主要限制了 LLMs 的输出结构，在 BSP 任务中，输出被限制为两个点或两个边界框，以及它们之间的相对位置关系。LLMs 的查询提示包含了需要进行推理的两个点或边界框，和对它们位置关系的询问。最终 LLMs 的输出结果如图 4.3（a）中红色箭头或蓝色箭头所示。

评估方法：根据预定义类别的不同，BSP 可以评估 LLMs 精确的方向感知和模糊的方向感知。BSP 被认为是分类任务，因此评估方法使用分类任务常用的评价指标，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 4.1　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。

$$P = \frac{TP}{TP + FP} \tag{4.5}$$

（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。

$$R = \frac{TP}{TP + FN} \tag{4.6}$$

（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。

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4.7}$$

4.3.2　页面位置预测

任务定义：在视觉富文档图像中，文本段（例如键-值对）通常存在多种固有的位置排列模式，例如上下、左右等。页面位置预测（Page Location Prediction，PLP）任务的目标是利用空间相邻字段的语义和整体布局信息来预测当前字段的位置。具体的说，受 Devlin 等人[73] 和 Xu 等人[5] 的启发，PLP 任务在测试文档中屏蔽了一定数量的字段位置，并要求 LLMs 根据文档上下文的语义和二维空间布局推断某个文本段缺失的位置坐标，即 “这个文本内容应该会出现在文档中的哪个位置?”。

图 4.4 展示了 PLP 任务的掩码过程和可能的预测结果。掩码指的是从原始的文档中选择一些文本段，保留文本的内容信息，然后把文本对应的边界框（即位置坐标）遮蔽，如图 4.4（b）所示，蓝色的序号和边界框代表被掩码的文本位置。（c）表示的是 ChatGPT 等 LLMs 通过上下文的文本与位置信息，预测这些位置掩码的文本应该出现的位置，如图中的红色序号和边界框所示。可以看到，预测的位置与真实的位置（蓝色）虽然有一定的差别，但在语义上是相近的。

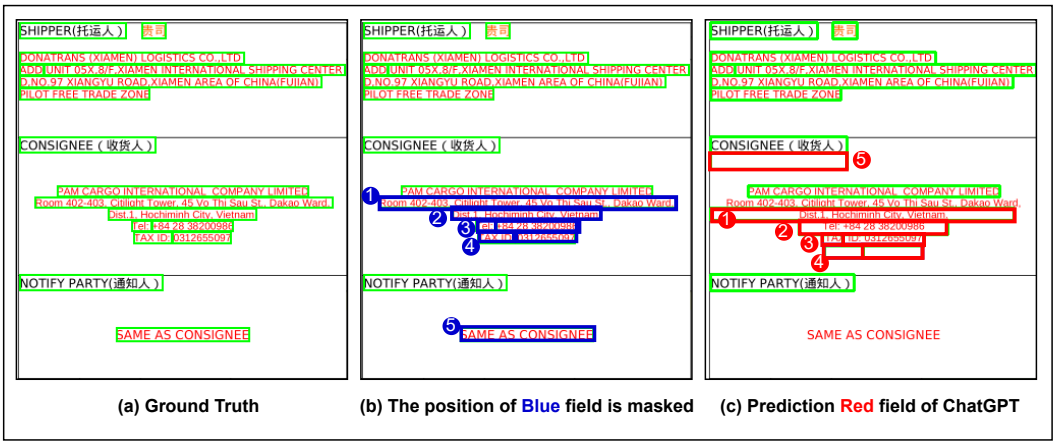


图 4.4　页面位置预测（PLP）任务的掩码和预测过程

任务流程：图 4.3（b）展示了 PLP 任务的主要流程。与 BSP 任务不同的是，PLP 任务被认为是一个位置回归的任务。在任务描述中需要告知 LLMs 具体的位置掩码策略，以及需要回归生成的位置形式。PLP 任务有一些必须要 LLMs 遵守的规则约束，例如预测的边界框的位置不能与文档中未被掩码的位置重合，因为需要预测的是一个新的文本段位置。输出格式被限制为一个被位置掩码的文本段和对应的边界框坐标。PLP 的查询提示包含了一个文档完整的 OCR 转录（除了被位置掩码的文本段外）以及对被掩码文本内容所在位置的询问。LLMs 最终的预测结果如图 4.3（b）中的红色边界框所示，预测的位置在真实位置的附近出现。

评估方法：PLP 任务预测的位置坐标无法通过精确的指标进行评测，因为即使预测的位置与真实位置没有重合，也不能认为 LLMs 的预测没有意义，因为大概率这些边界框的语义位置是合理的。因此，PLP 通过对掩码位置的预测偏离程度来衡量 LLMs 位置生成的准确度。由于 LLMs 预测的矩形边界框之间的距离难以进行衡量，PLP 的评估针对于边界框的中心位置坐标计算偏差。常用的空

间距离度量有欧氏距离（Euclidean Distance）、曼哈顿距离（Manhattan Distance）和切比雪夫距离（Chebyshev distance）。三种距离评估方式的区别如图 4.5 所示，计算公式如下：

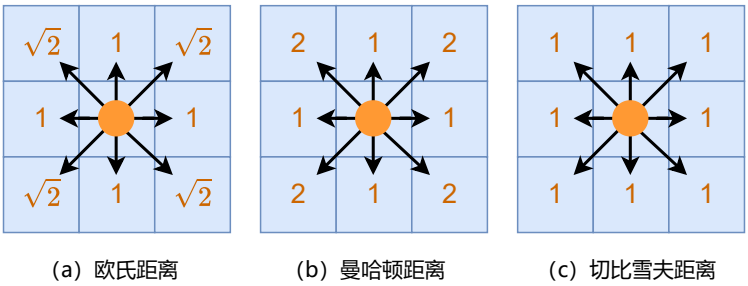


图 4.5 三种空间距离度量的比较

（1）欧氏距离 D_{eu}，即两点之间的最短距离是连接两点的直线距离，以二维空间计算为例：

$$D_{eu} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4.8}$$

（2）曼哈顿距离 D_{ma}，不是通过计算两点间的直线距离来确定两点的距离，而是通过将这两点在各坐标轴上的投影长度进行累加得出：

$$D_{ma} = |x_1 - x_2| + |y_1 - y_2| \tag{4.9}$$

（3）切比雪夫距离 D_{ch}，将两点之间的距离定义为其各坐标数值差的最大值：

$$D_{ch} = \max(|x_2 - x_1|, |y_2 - y_1|) \tag{4.10}$$

4.3.3 文档信息抽取

任务定义：文档信息抽取（Document Information Extraction，DIE）任务通常包含语义实体识别（Semantic Entity Recognition，SER）和关系抽取（Relation Extraction，RE）。SER 任务的目标是要求 LLMs 理解文档中实体语义和位置关系，并且给该实体推断一个属于它的类型。在 LLMs 的二维空间位置理解评估中，SER 这个任务包含粗粒度 SER 和细粒度 SER。粗粒度的 SER 只关心语义实体的键和值的识别，而细粒度的 SER 需要预测语义实体的具体类别。此外，RE

任务在细粒度 SER 的基础上，不仅要区分语义实体的键和值，还要对键和值的具体类别进行预测。

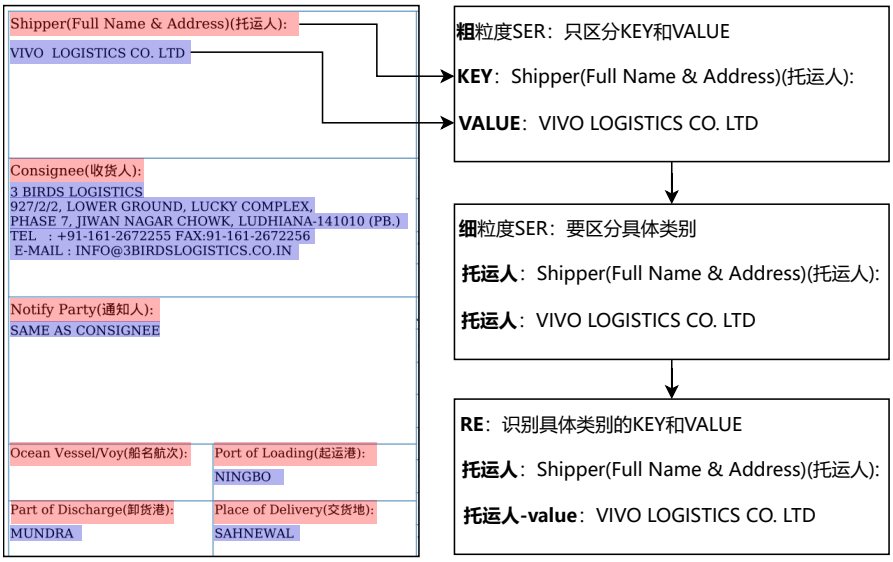


图 4.6　DIE 任务粗细粒度语义实体识别 SER 和关系理解 RE 示例

图 4.6 展示了粗粒度语义实体识别（SER）、细粒度语义实体识别 SER 和 RE 任务的对比。以视觉富文档为例，粗粒度语义实体识别 SER 主要关心的类别是键（key）和值（value），细粒度的 SER 主要关心的是某一具体的类别，如发货人、起运港、毛重等等，而 RE 任务更关注键和值之间的匹配关系，例如发货人和发货人-value，这是一种更细粒度的类别区分，包含了 LLMs 对实体间的关系理解。

任务流程：图 4.3（c）展示了 DIE 任务的主要流程。任务描述中需要提示 LLMs 目前执行的是粗粒度 SER、细粒度 SER 还是 RE 任务，并且更新相应的标签映射集合。DIE 任务被认为是序列标注任务，所以输出格式中需要指定序列号，即语义实体 ID，除此之外，还要有实体的文本内容、边界框位置和 LLMs 预测的类型。DIE 任务的查询提示包含了整个文档 OCR 转录的信息，以及对每个实体类别的询问。LLMs 执行的 SER 任务的预测结果如图 4.3（c）中的橙色边界框所示，而图中蓝色和红色边界框分别代表了 RE 任务的预测结果，即键值对的关系。

评估方法：DIE 任务被视作序列标注任务，这是一种多分类的问题。与 BSP 的评估方法类似，精确度、召回率和 F1 分数作为 LLMs 文档信息抽取任务的性

能衡量。除此之外，DIE 任务中涉及到许多特定的类别，这些类别在大量的文档中出现了明显的长尾分布，即有实体数量较多的类别和较少的类别，这些类别的预测结果更能展示 LLMs 对于二维空间位置理解的能力。因此，每个细分类别的精确度、召回率和 F1 分数也是 DIE 任务评估的重点。

4.3.4 文档布局生成

任务定义：由于 DIE 主要是局部键值关系的理解，对文档整体排版布局的依赖程度较低。而文档布局生成（Document Layout Generation，DLG）则更依赖文档全局的排版信息，要求 LLMs 生成符合具体类别先验概率分布的布局结构。具体的说，本文中 DLG 任务被认为是离散的二维随机变量的概率分布问题，以具体类别为例，“发货人”这个类别通常出现在文档的左上角，即在文档左上角出现的概率更大。为了更明确的定义“发货人”的概率分布，如果此类别文本段的左上角 x 坐标落在（0，100），y 坐标落在（0，100）的区间内，那么就认为该文本段出现在（100，100）这个坐标点，因此类别 c 的离散二维概率分布可以形式化为：

$$\begin{aligned} &P^c(x_i - inter < X \leq x_i, y_j - inter < Y \leq y_j) \\ &= P^c(X = x_i, Y = y_j) = p^c_{ij} = \frac{count^c_{ij}}{count^c_{sum}} \end{aligned} \tag{4.11}$$

其中 $inter$ 指的是区间间隔，按照上面的例子来说，取值为 100。$count^c_{ij}$ 指的是 c 类别的文本段落在区间内的个数，$count^c_{sum}$ 指的是 c 类别文本段的总个数。

为了更细致的刻画具体类别的分布情况，DLG 任务将二维随机变量的概率分布分解为两个子分布，即 x 方向和 y 方向的一维概率分布，分别被形式化为：

（1）x 方向的概率分布

$$P^c_x(x_i - inter < X \leq x_i) = P^c_x(X = x_i) = p^c_i \tag{4.12}$$

（1）y 方向的概率分布

$$P^c_y(y_j - inter < Y \leq y_j) = P^c_y(Y = y_j) = p^c_j \tag{4.13}$$

这两个分布更直观的展示了文档中关键信息所处的位置情况，根据真实情

况和预测的分布一致性差异可以分析 LLMs 生成布局结构的合理性。

任务流程：图 4.3（d）展示了 DLG 任务的主要流程。任务描述中需要提示 LLMs 需要利用文档全局的信息和本身所了解的先验知识，如海运单、收据等大致版式结构和关键信息排列，以便 LLMs 准确的使用这些知识。为了更好的限制 LLMs 文档布局生成的结果，DLG 任务提供了先验规则，优化模型的输出。输出格式被限制为边界框位置和所属的类别。DLG 的查询提示主要有文档的 OCR 转录和标注的类型，以及一个对文档版式布局生成的询问。LLMs 最终的输出是与原始文档布局相似的新布局，并且符合一定先验知识。图 4.3（d）中的橙色边界框是 LLMs 预测的“发货人”类别的文本段的大致分布位置。

评估方法：DLG 任务的评估涉及到 LLMs 预测的概率分布和真实分布之间的比较。由第 三 章可知，衡量概率分布之间的差异的评价指标通常有 KL 散度[144]和 JS 散度[143]，计算公式如下：

（1）KL 散度（Kullback-Leibler 散度）：KL 散度衡量的是从一个概率分布 Q 到另一个概率分布 P 的不匹配程度，是一种单向的过程。

$$D_{KL}(P\|Q) = \sum_{i} P(i) \log\left(\frac{P(i)}{Q(i)}\right) \tag{4.14}$$

（2）JS 散度（Jensen-Shannon 散度）：JS 散度是对 KL 散度的一种改进，它衡量的是两个概率分布之间的相似性，是一种双向的过程。JS 散度通过计算两个分布的平均分布与每个分布之间的 KL 散度来定义。它的公式如下：

$$D_{JS}(P\|Q) = \frac{1}{2}\left(D_{KL}(P\|M) + D_{KL}(Q\|M)\right) \tag{4.15}$$

4.4 实验

本小节主要介绍 LLMs 二维空间理解的评估数据集及构建方法、四个评估方案的实验结果与可视化分析，以及在这基础上提出的基于布局提示的 LLMs 文档实体关系理解方法。

4.4.1 数据集介绍

为了简单客观的进行四个二维空间理解评估任务，本章基于第三章标注的 SEABILL 数据集构建了四个数据子集，分别为 BSPData，PLPData，DIEData 和

DLGData。下面分别介绍这四个数据集的信息、构建过程和数据形式：

（1）BSPData：BSPData 数据子集由 466 张海运单图片和对应的 466 组（group）边界框集合组成。每一组边界框集合都是从对应 SEABILL 数据文档的 OCR 转录中随机采样的，包含 10 个随机边界框对（bbox pairs）。这样采样的目的是保留了每个文档中边界框的实际位置来帮助 LLMs 理解后续更有挑战的任务。每一对边界框的黄金相对位置关系，如算法 1 一样计算，包括了上、下、左、右、左上、左下、右上和右下八种方向的感知。BSPData 子集按需求被划分成了提示验证集和测试集，其中提示验证集 200 组，测试集 266 组。BSPData 数据子集以 JSON 的文件形式存放。

（2）PLPData：由于 LLMs 的输入长度限制，PLP 实验从 SEABILL 数据集中采样了文本段数少于 50 的 466 份文档进行数据子集的构建。PLP 任务要求给定的数据集中包含被掩码的类别字段，如掩码“发货人”这个类别的边界框位置，以供 LLMs 预测。因此，可以根据海运单中的关键信息，如发货人（key）、发货人-value（value）、唛头[①]（key）、唛头-value（value）等类型来构建数据子集。以掩码“发货人”这个类型为例，具体的数据子集构建流程如图 4.7 所示。

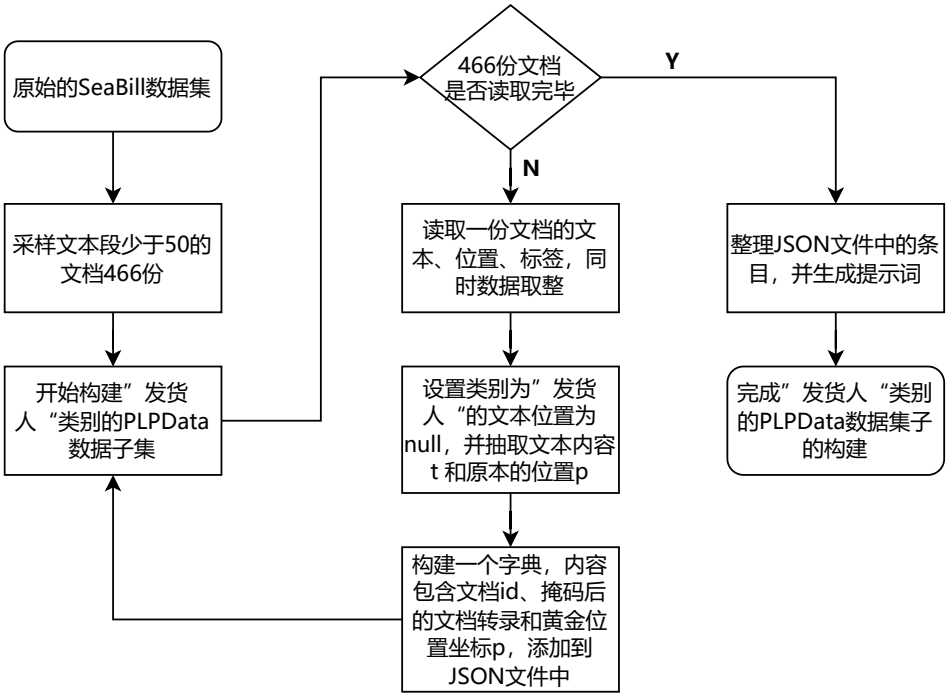


图 4.7 PLPData 数据子集的构建流程

① 唛头一般指商标，用文字、图形和记号标明在货物的包装上，以便识别一批货物不同于另一批货物的标记。

（3）DIEData：与 PLPData 一致，受到输入长度的限制，DIEData 子集也是从 SEABILL 数据集中采样得到的，包含了 466 份文档图像和对应的文本、位置和类别标注。对于粗粒度的 SER 任务，DIE 任务预定义了 {Key、Value、Other} 三个类别，让 LLMs 只关注实体的键值属性；对于细粒度的 SER 任务，预定了 {发货人，收货人，起运港，目的港，唛头，毛重...} 等 56 个类别；而对于 RE 任务，需要 LLMs 在细粒度类别上进一步区分发货人（key）和发货人-value（value）。因此，在标注时，每一个文本段都被标注为一个最细粒度的键值实体，方便后续的实验。

（4）DLGData：DLGData 子集需要评估 LLMs 每个类型关键信息的分布情况。与 PLPData 子集的构建方法类似，DLG 任务将 466 份文档中的关键字段按照类型划分成组。以“发货人”类型为例，将所有文档中的“发货人”类型的语义实体收集起来，每一项包含文本内容和边界框位置。这样，真实数据中所有“发货人”类型的实体可以在 1000*1000 或者 1*1 的文档页面中展示。散点分布被认为是“发货人”类型的黄金概率分布，可以通过公式 4.11 计算。

4.4.2 实验环境与细节

本章实验选用的 LLMs 及其版本的统计信息如表 4.2 所示。

表 4.2 选用的 LLMs 的参数信息统计

模型名称	发布时间	内核版本	参数规模	训练数据量	上下文长度	语言	是否开源
CogVLM	2023-10	视觉：EVA2-CLIP-E 语言：Vicuna-17B-v1.5	17B	15 亿张图文对	8K	英	✔
ChatGPT	2022-11	gpt3.5-turbo	未知	近 1 万亿个单词	16K	中、英	✖
GPT-4	2023-03	gpt4-0125-preview	未知	近 13 万亿个单词	8K	中、英	✖

4.4.3 评估实验结果与分析

（1）基本空间感知（BSP）

为了评估 LLMs 最基本的局部空间感知能力，BSP 方案设置了两种不同难度的实验，包括点对点（point-to-point）的相对位置理解和边界框对边界框（bbox-to-bbox）的相对位置理解。其中，点对点的实验相对更加容易，因为模型只需要在确定的坐标系中判断两个点的位置关系，而边界框的实验则需要考虑更多的

方向区域，这是视觉富文档理解的必要能力。

由于基于 LLMs 的实验结果严重依赖于提示构建的质量，因此，本章中的所有任务在进行评估之前都会先进行 LLMs 提示的验证。这个过程包含候选提示构建、提示评估和提示优化三个步骤，并且可能进行多轮的迭代，直到选出最合适的提示进行后续的任务评估。注意，每个任务的提示一旦选定，将不会再进行调整。具体的验证过程如下：

- **候选提示构建**：针对每个任务预先构建三个候选的提示，候选提示的构建遵循如图 3.2（b）所示的模板，每个提示会微调其中的若干部分内容或调整内容的顺序。以 BSP 任务为例，候选提示间主要的区别在于对位置判断算法的解释。
- **提示评估**：提示评估旨在挑选质量最好的候选提示，这个过程基于各个数据子集中划分的验证集，验证集和测试集以 1：1 的比例进行划分。LLMs 将不同的候选提示作为输入，得到验证的结果。候选提示根据验证的结果好坏进行排序，最后选取结果最好的提示作为后续提示优化的目标。
- **提示优化**：提示优化的目标是让模型自己理解提示的内容。检查 LLMs 对提示中每一部分的理解，可以知道构建的提示是否被完全理解，或者有哪个部分难以被正确理解。通过这个过程，可以确保 LLMs 完全理解提示的语义信息，进而进行下一次的迭代验证。

经过验证后，作为 LLMs 输入的提示大幅增强了模型的泛化能力，避免了在评估过程中由于提示质量不一致导致的高偏置和高方差。

表 4.3　LLMs 零样本点的相对位置预测的评估结果

Model（point-to-point）	**精确方向**			**模糊方向**		
	P % ↑	**R %** ↑	**F1 %** ↑	**P %** ↑	**R %** ↑	**F1 %** ↑
ChatGPT(gpt3.5-turbo)	58.99	62.33	60.61	98.64	92.27	95.35
GPT4(gpt4-0125-preview)	**99.33**	**98.40**	**98.86**	**99.53**	**98.60**	**99.06**

本实验基于 BSPData 数据子集，在 400 份评估数据上进行实验。BSP 任务除了关心 LLMs 精确方向（精准判断唯一的方向）的判别能力，还进行了模糊方向（主要方向）的比较。模糊方向的定义为两个目标之间的大致方位，例如某两个 bbox 之间的黄金位置关系为 “左上”，那么 LLMs 预测位置关系为 “左” “上” “左

表 4.4 LLMs 零样本边界框的相对位置预测的评估结果

Model（bbox-to-bbox）	精确方向			模糊方向		
	P % ↑	R % ↑	F1 % ↑	P % ↑	R % ↑	F1 % ↑
ChatGPT(gpt3.5-turbo)	65.72	43.47	52.33	89.55	**68.67**	**77.73**
GPT4(gpt4-0125-preview)	**87.91**	**62.70**	**73.20**	**89.90**	64.10	74.84

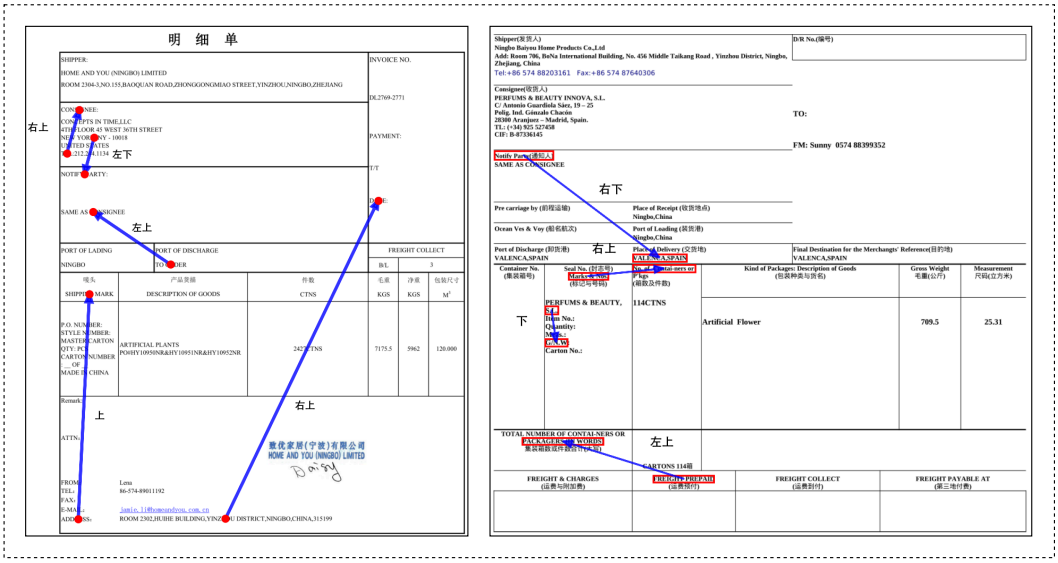


图 4.8 ChatGPT 在 BSP 评估任务上的可视化案例

上”这三种关系都被认为是正确的。因此，模糊方向的实验设置较好的评估了 LLMs 粗粒度的方向感知，而精确方向的实验评估 LLMs 更细粒度的方向感知。

表 4.3 和表 4.4 分别展示了 LLMs 零样本点的相对位置预测的评估结果和边界框的相对位置预测结果。如表 4.3 所示，在点的实验中，ChatGPT 对于点的精确位置判断得分为 60.61%，与 GPT4 之间的差距接近 40%，但是在模糊方向的判断中，两模型之间的差距缩小到 5% 以内。此外，实验结果展示了 GPT4 几乎可以完全理解点之间二维位置，而 ChatGPT 更擅长模糊方向的感知。在 VRDs 更关心的边界框的实验中，如表 4.4 所示，所有模型的评估结果相较于点的实验，都有显著的下降，这说明边界框的位置关系判断更复杂，因为要额外考虑区域的问题。与点实验一致，所有模型对于边界框模糊方向的感知相比精确方向都有不同程度的提升。令人惊讶的是，ChatGPT 的模糊方向感知能力甚至比 GPT4 更强。总之，BSP 的评估展示了文本预训练的 LLMs 具有强大二维感知潜力，GPT4 对于精确和模糊的方向判断能力都很优秀，而 ChatGPT 更适合模糊方向的判断。

图 4.8 展示了 ChatGPT 在 BSP 评估任务上的可视化案例，左图为点的实验，

右图为边界框的实验。蓝色箭头的方向为第二个点或边界框相对于第一个点或边界框的相对位置方向。可视化案例更直观的展示了 LLMs 的基本空间感知能力。其中左图展示了点与点（point-to-point）的相对位置关系，从图中可以看出 ChatGPT 可以较为精确的判断文档中两个随机位置点的方向关系。右图则展示了边界框和边界框（bbox-to-bbox）之间的位置关系，ChatGPT 依据给定的判断标准合理的预测了不同边界框之间的位置关系，这是视觉富文档实体语义和关系理解的重要基础。

（2）页面位置预测（PLP）

为了评估 LLMs 在视觉富文档上的语义实体位置理解能力，PLP 评估任务从不同的语义类别出发，计算 LLMs 预测位置与真实黄金位置之间的一致性分数，通过该分数来衡量模型能力的好坏。

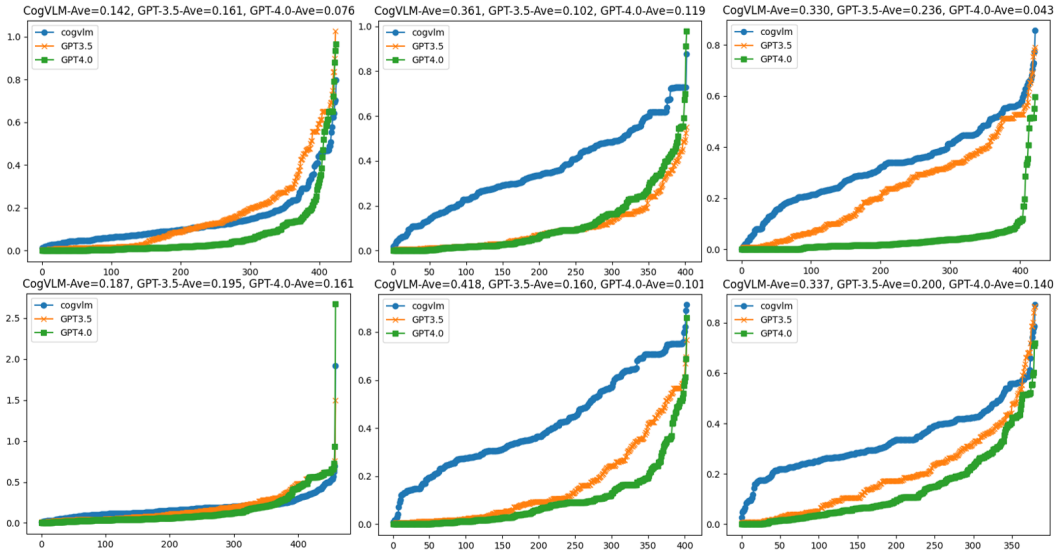


图 4.9 LLMs 预测不同类别的边界框位置的偏差（基于欧式距离）

评估实验基于 PLPData 数据子集，该数据集包含了 56 组按类别划分的位置掩码数据。本实验主要统计了“发货人”、“发货人-value”、“起运港”、“起运港-value”、“唛头”和“唛头-value”这六个类别的位置生成结果，并基于欧氏距离计算位置一致性分数，分数越高代表 LLMs 的预测结果越接近真实的黄金位置，进一步表明 LLMs 具有更强的实体位置理解能力。图 4.9 展示了 LLMs 预测不同类别的边界框位置的偏差情况，距离由小到大排序。第一列的结果是“发货人、发货人-value”，第二列的结果是“起运港、起运港-value”，第三列的结果是

表 4.5 LLMs 预测不同类别的边界框位置的一致性分数

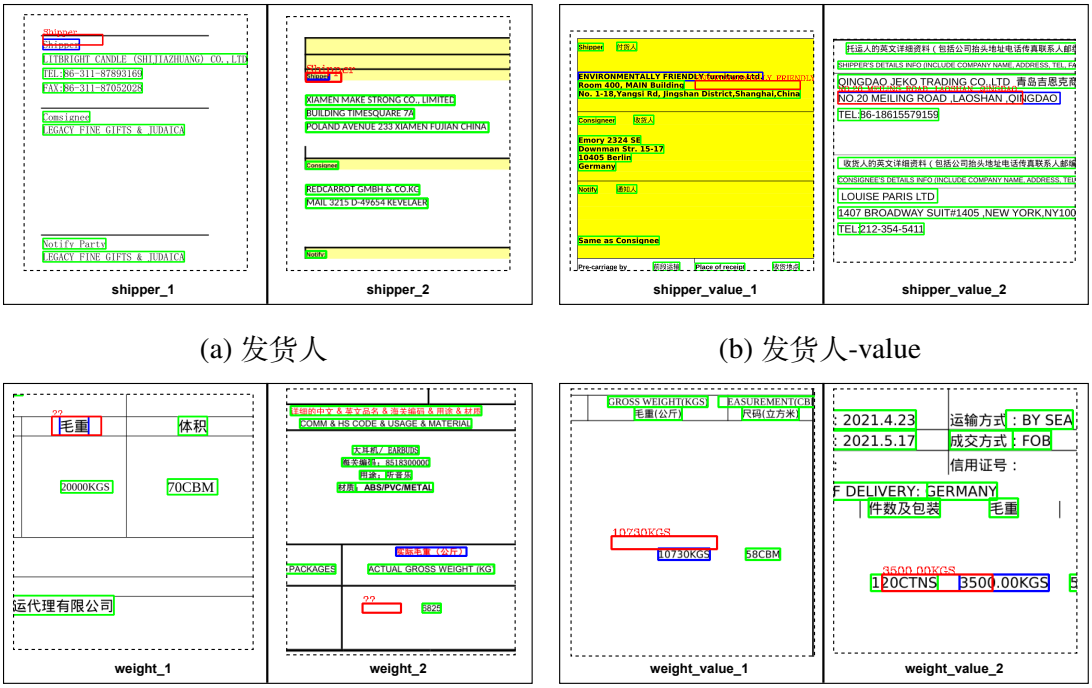
Model	位置一致性分数% ↑						
	发货人	发货人-v	起运港	起运港-v	唛头	唛头-v	Average
CogVLM	85.8	81.3	63.9	58.2	67.0	66.3	70.4
ChatGPT(gpt3.5-turbo)	83.9	80.5	**89.8**	84.0	76.4	80.0	82.4
GPT4(gpt4-0125-preview)	**92.4**	**83.9**	88.1	**89.9**	**95.7**	**86.0**	**89.3**

“唛头、唛头-value”。图中的横坐标指的是样本编号，纵坐标指的是欧式距离偏差。为了更好的展示距离偏差的趋势，该实验的样本编号依据偏差距离由小到大排序。有意思的一点是，图中距离偏差的曲线与横轴所围图形的面积可以被认为是该模型在所有样本数据上的总偏差，即面积越小，直观的表示为模型位置生成的偏差越小。从图中可以看出，较小的视觉语言模型 CogVLM 在“发货人”和“发货人-value”类别上的预测效果十分优秀，赶上了一众的超大语言模型，但在其他类别上明显弱于 GPT 系列的模型。此外，GPT4 在所有类别上都取得了极好的分数。结合表 4.5，该表定量的展示了 LLMs 预测不同类别的边界框位置的一致性分数，分数越高表明 LLMs 预测的位置越精准。粗体表示最好的结果。“-v”的后缀表示类别所对应的值。从平均值的角度看，GPT4 超 ChatGPT7 个百分点，并远超了 CogVLM 接近 30 个百分点。这一方面说明了 GPT 系列的超大语言模型在提示的引导下，已经可以较好的理解视觉富文档中的语义实体位置。另一方面说明小型视觉语言模型也具有强大的潜力去解决视觉富文档的问题。

图 4.10 展示了 ChatGPT 在 PLP 评估任务上的可视化案例，蓝色边界框代表真实情况，红色代表预测结果，绿色边界框代表文档中的其余实体。可视化案例更直观的展示了 ChatGPT 的实体位置预测能力。从图中可以发现，ChatGPT 预测的边界框位置大致有两种情况：第一种，如“发货人”类别（a）所示，ChatGPT 预测的结果几乎覆盖了大部分的真实位置，所在的位置十分精准。第二种，如“毛重-value”类别（d）所示，ChatGPT 预测的边界框位置虽然有一些偏差，但都分布在真实边界框的附近，在语义的角度上模型预测的结果是合理的。总之，LLMs 生成的这些位置包含了语义的理解和模型自身的想象力，这对视觉富文档的理解和生成任务具有重要的意义。

（3）文档信息抽取（DIE）

为了评估 LLMs 在视觉富文档上的语义实体关系理解能力，DIE 评估任务包



(a) 发货人　(b) 发货人-value

(c) 毛重　(d) 毛重-value

图 4.10　ChatGPT 在掩码不同类别上的预测位置

含三种不同粒度的实体抽取任务，包括了粗粒度语义实体识别（粗粒度 SER），细粒度语义实体识别（细粒度 SER）和语义实体关系理解（RE），这三种实验设置的难度逐渐递增。其中，粗粒度 SER 仅区分实体的键值属性，细粒度 SER 需要识别特定的类别，而 RE 结合前两种方案，进一步识别具体类别的键值属性。信息抽取是视觉富文档理解的核心方法。

表 4.6　LLMs 零样本粗粒度语义实体识别 SER 的评估结果

Model（粗粒度 SER）	**Average**			**ID（域内）**			**OD（域外）**		
	P % ↑	**R % ↑**	**F1 % ↑**	**P % ↑**	**R % ↑**	**F1 % ↑**	**P % ↑**	**R % ↑**	**F1 % ↑**
ChatGPT(gpt3.5-turbo)	55.17	54.68	54.93	79.31	48.93	60.52	79.31	48.93	60.52
GPT4(gpt4-0125-preview)	**67.54**	**57.03**	**61.84**	**91.18**	**64.83**	**75.78**	-	-	-

表 4.7　LLMs 零样本细粒度语义实体识别 SER 的评估结果

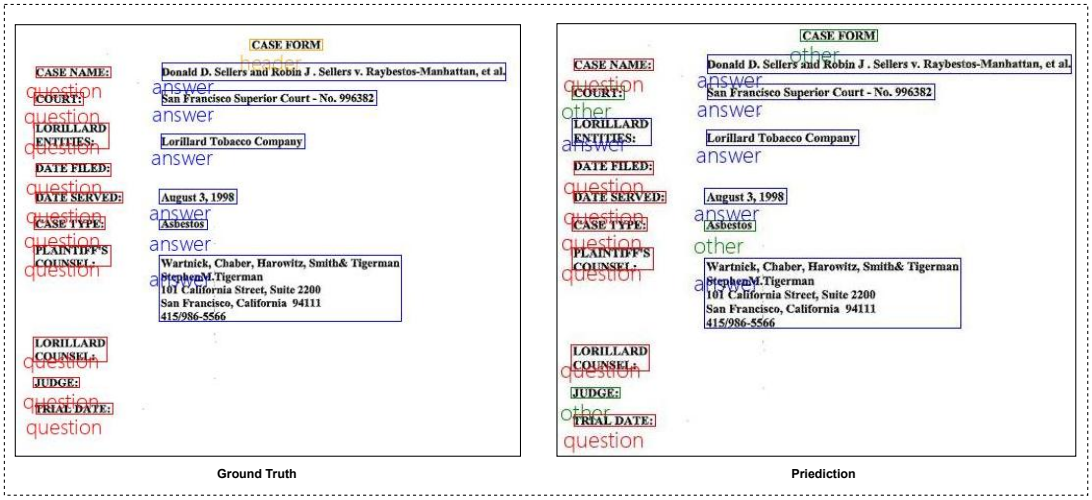
Model（细粒度 SER）	**Average**			**ID（域内）**			**OD（域外）**		
	P % ↑	**R % ↑**	**F1 % ↑**	**P % ↑**	**R % ↑**	**F1 % ↑**	**P % ↑**	**R % ↑**	**F1 % ↑**
ChatGPT(gpt3.5-turbo)	60.61	60.07	60.34	90.92	57.28	70.28	59.17	57.84	58.50
GPT4(gpt4-0125-preview)	**68.31**	**68.06**	**68.19**	**92.70**	**77.92**	**84.67**	-	-	-

表 4.8 LLMs 零样本关系理解 RE 的评估结果

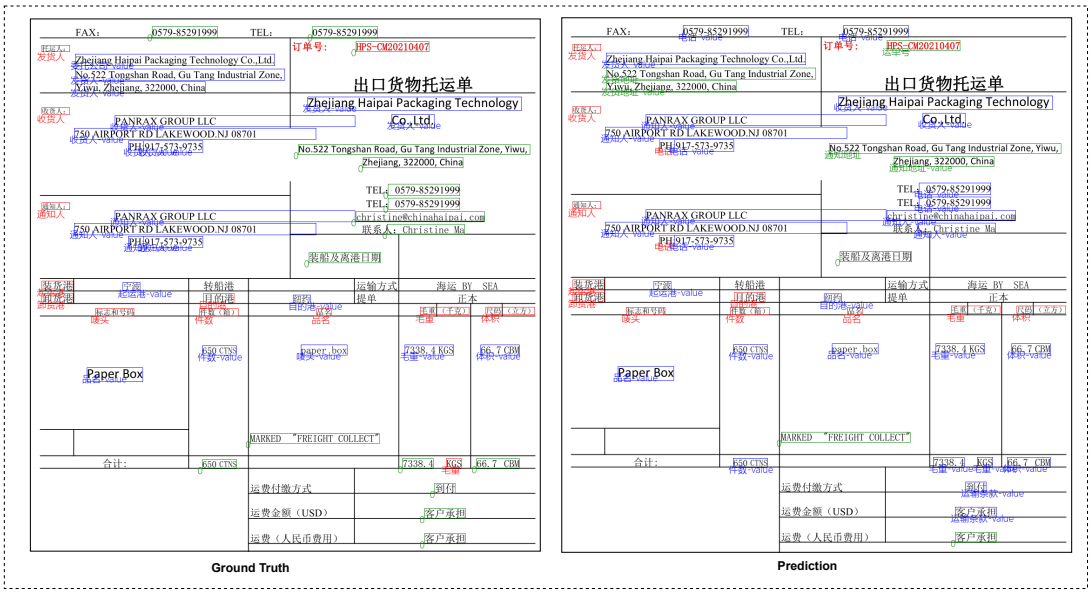
Model（RE）	Average			ID（域内）			OD（域外）		
	P % ↑	R % ↑	F1 % ↑	P % ↑	R % ↑	F1 % ↑	P % ↑	R % ↑	F1 % ↑
ChatGPT(gpt3.5-turbo)	53.63	53.16	53.39	76.00	47.88	58.75	52.52	51.33	51.92
GPT4(gpt4-0125-preview)	**65.76**	**65.52**	**65.64**	**88.58**	**74.46**	**80.91**	-	-	-

DIE 实验基于 DIEData 数据子集，该数据集包含了 466 份带有类别标注的 SEABILL 数据。对于每一个实体抽取任务，评估实验除了关心总体类别的平均实验结果，还进行了域内（ID）和域外（OD）的比较，域内指的是仅让 LLMs 评估黄金标注不为“Other”的类别，保证 LLMs 的预测结果和黄金标注之间的可比性。域外的评估只让 LLMs 评估黄金标注为“Other”的类别，由于黄金标注中的这些类别没有进一步的细分，所以本实验将 GPT4 的预测结果作为黄金标注来评估其余的 LLMs。采用域内域外的实验设置旨在更客观的评估 LLMs 的理解能力，解决 LLMs 在分类任务中的评估问题。这个评估问题的解释可见图 4.11 (b)，以图中右上角的“订单号”文本为例，在黄金标注中，此实体的类别是不被关心的类别，因此被标注为“Other”（绿色），但是 LLMs 具有强大的先验知识，可以很好判断此实体的类别为“运单号”，这会导致在直接进行评估时，LLMs 的性能受到较大的限制。本实验的域外评估将这些不关心的类别单独进行评价，解决这一评估问题。

表 4.6、表 4.7 和表 4.8 分别展示了 LLMs 在零样本条件下粗细粒度 SER 任务和关系理解 RE 任务的评估结果。在细粒度 SER 的实验中，LLMs 域内的结果都比平均结果高 10 个百分点以上，特别的，GPT4 的结果更是高了 15 个百分点，这说明 DIE 任务的域内实验可以进一步客观的评估 LLMs 的能力。此外，域外的实验结果展示了通过 LLMs 的先验知识，可以将数据集中 58.5% 不关心的实体识别成新的类别，这是 LLMs 相较于传统多模态文档智能模型的一大优势。与细粒度 SER 任务相比，LLMs 的 RE 实验结果在平均，域内和域外都有一定程度的下降，这是可以理解的，因为关系理解任务可以被认为是在细粒度 SER 任务的基础上进行键值抽取。具有相同规律的是，LLMs 的域内结果相比平均都有显著的提升。总之，从实验结果来看，GPT4 的 VRDs 信息抽取能力，尤其是关系理解是远超 ChatGPT 及其他模型的。此外，ChatGPT 的零样本表现也是具有潜力的，在引入上下文范例和相应的布局提示，ChatGPT 或更小的 LLMs 有望很好



(a) 粗粒度语义实体识别（SER）



(b) Key-value 关系理解（RE）

图 4.11　零样本条件下 ChatGPT 粗粒度 SER 和关系理解 RE 的可视化案例

的解决 VRDs 的信息抽取任务。

图 4.11 展示了在零样本条件下 ChatGPT 粗粒度 SER 和 RE 关系理解的可视化案例，其中红色边界框代表类别的“键（key）”，蓝色边界框代表类别的“值（value）”，绿色边界框代表文档中的其余实体。可视化案例更直观的展示了 ChatGPT 的实体关系识别能力。从粗粒度 SER（a）结果中可以看出，ChatGPT 能很好的根据语义区分实体的键值属性，虽然有“CASE FROM”等的实体识别错误，可能的原因是在零样本的前提下，LLMs 的先验知识主导了语义的识别，忽视了较为重要排版布局的信息。此外，RE 关系理解（b）的实验结果是有趣的，

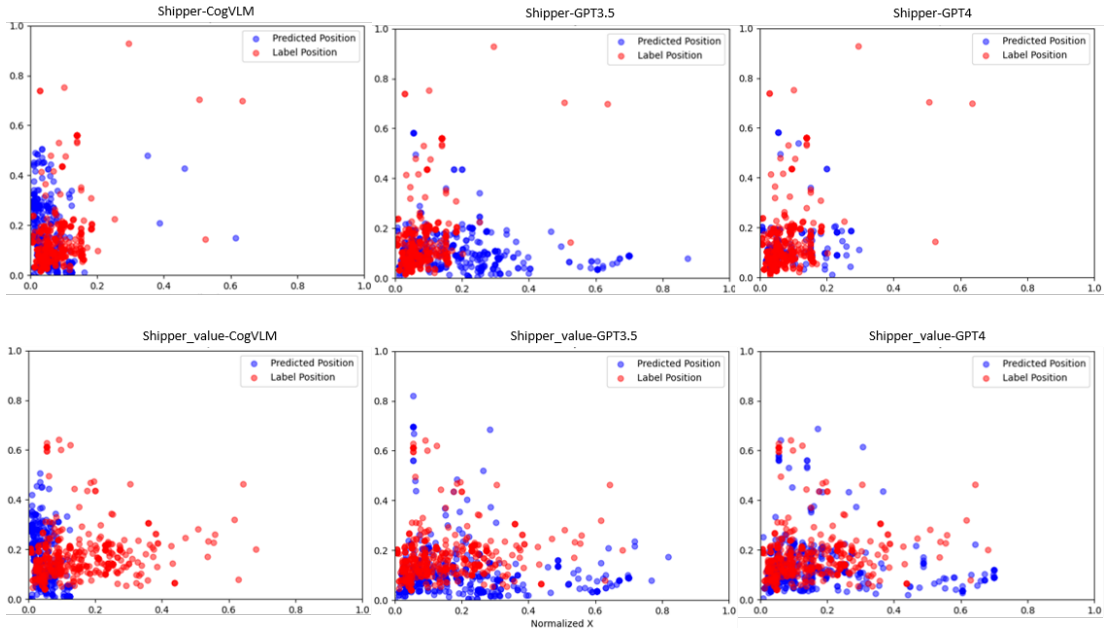


图 4.12　LLMs 预测“发货人”和“发货人-value”的实体位置的散点分布

LLMs 不仅有能力将预定义的实体类型识别正确，还可以将一些“Other”（绿色）类别的实体识别成合理的新类别，例如传真号码和电话号码，LLMs 预测为“电话-value”的新类别，这是合理的。这从另一个角度证明了 LLMs 丰富的先验知识和基本的空间感知能力，可以有效的帮助 LLMs 对视觉富文档的理解，并且 LLMs 的优势是不会局限于给定的类型，具有强大的零样本知识迁移和创造力。

（4）文档布局生成（DLG）

为了评估 LLMs 对视觉富文档的实体布局生成能力，DLG 评估任务主要围绕二维和一维随机变量的概率分布来衡量 LLMs 的实体布局生成能力。其中二维随机变量分布被形式化为在 1*1 的二维空间中的散点分布情况，而一维概率分布从 x 和 y 两个不同的方向描述布局生成的概率分布。

DLG 实验基于 DLGData 数据子集，该数据集包含 56 组按类别划分的文档实体，这是黄金的先验分布的来源。本实验主要统计了“发货人”、“发货人-value”、“起运港”、“起运港-value”、“唛头”和“唛头-value”这六个类别的生成概率分布，并基于 KL 散度和 JS 散度与黄金先验分布计算一致性分数，分数越高代表 LLMs 的布局生成结果越接近于真实的实体分布情况，进一步表明 LLMs 具有更强的实体布局生成能力。

图 4.12 、图 4.13 和图 4.14 分别展示了 LLMs（CogVLM，GPT3.5，GPT4）

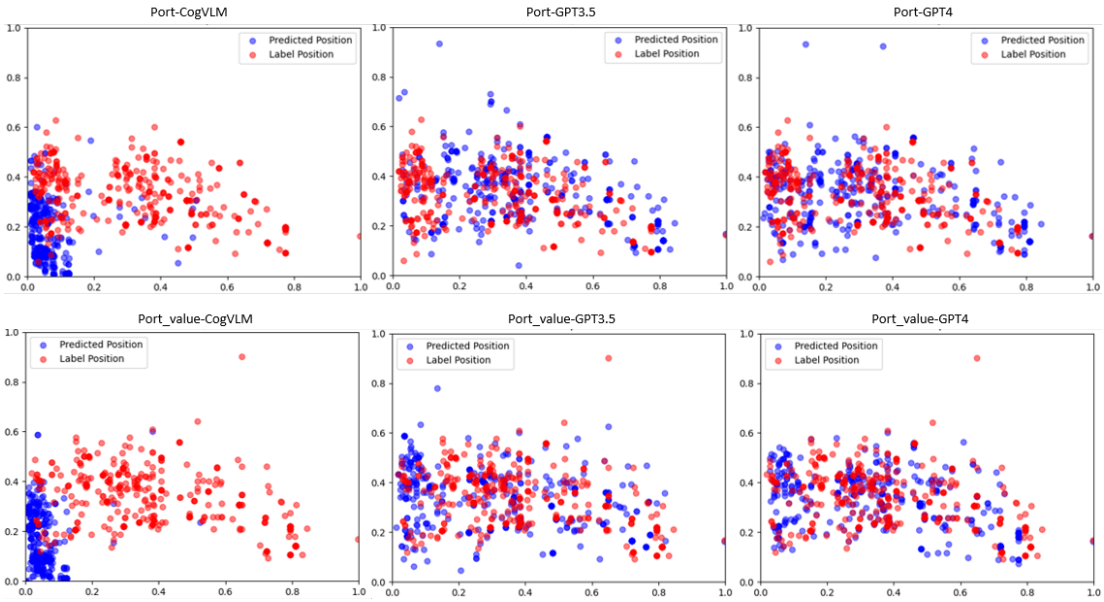


图 4.13　LLMs 预测“起运港”和“起运港-value”的实体位置的散点分布

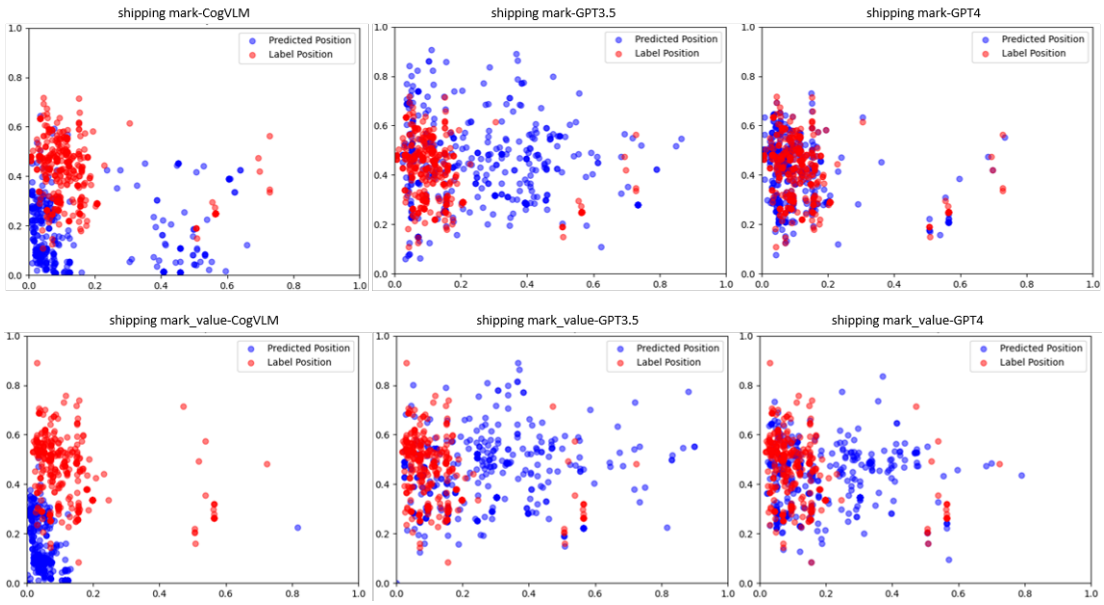


图 4.14　LLMs 预测“唛头”和“唛头-value”的实体位置的散点分布

预测发货人系列、起运港系列和唛头系列实体位置的散点分布图。上面的图展示的是“键（key）”类别的预测分布和真实分布，下面的图是对应的“值（value）”类别的预测分布和真实分布情况。从直观上看，CogVLM 在“发货人”和“发货人-value”的位置分布预测上表现良好，但在其他类别的预测中显示出较大的局限性，而 ChatGPT（GPT3.5）和 GPT4 整体上预测的效果都较为良好，图中蓝色的散点与红色的散点有大面积重合或相似的分布。总之，散点图的结果说明了 LLMs 合理利用先验知识可以在文档布局生成任务上展现极大的潜力。

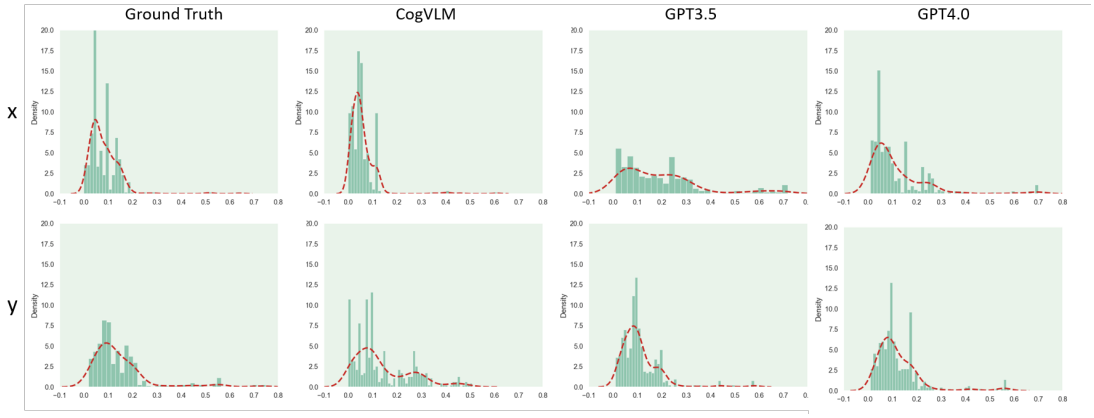


图 4.15　LLMs 预测“发货人”类别的实体位置在 x 方向和 y 方向的概率分布

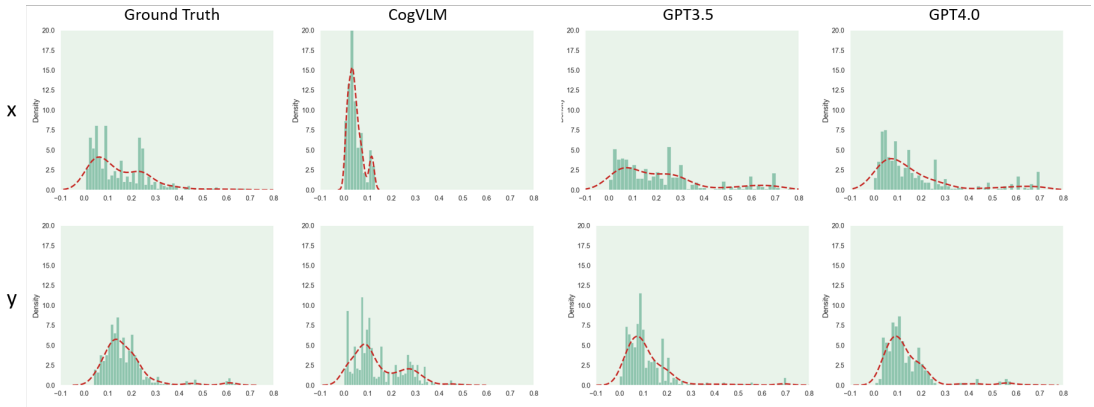


图 4.16　LLMs 预测“发货人-value”类别的实体位置在 x 方向和 y 方向的概率分布

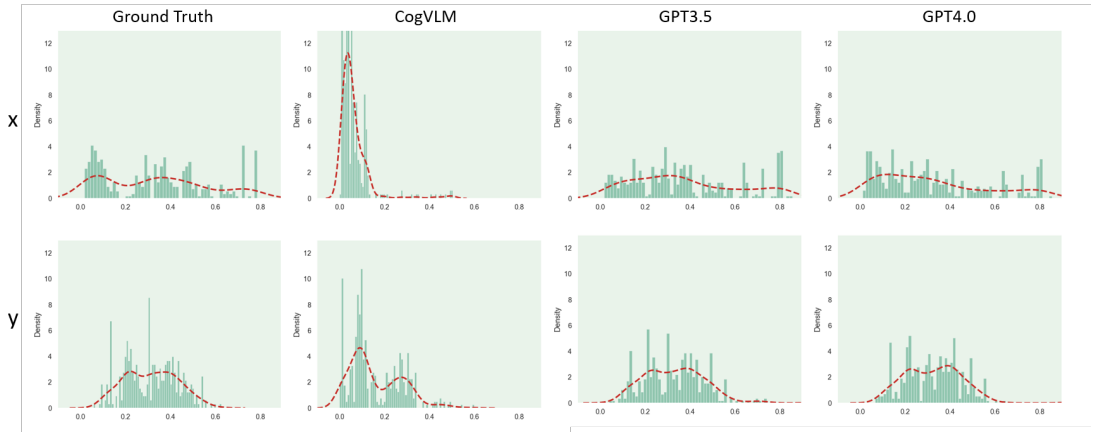


图 4.17　LLMs 预测“起运港”类别的实体位置在 x 方向和 y 方向的概率分布

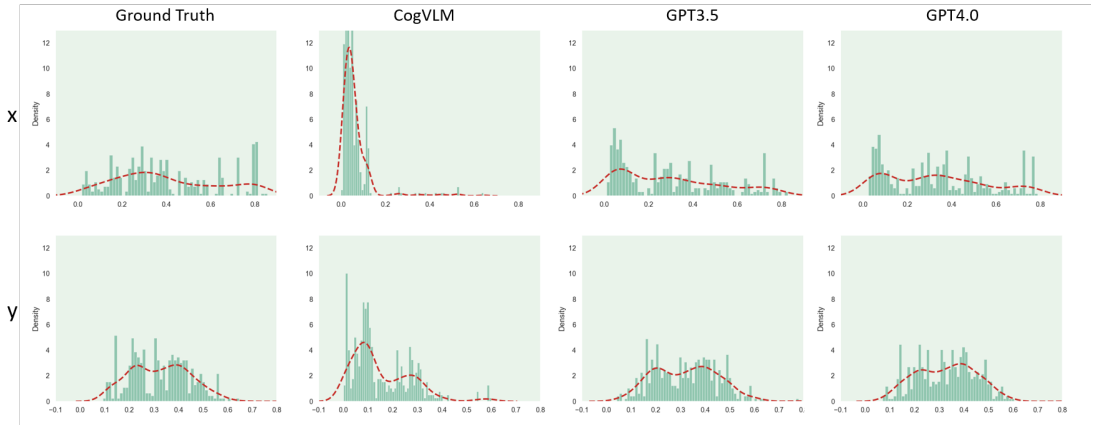


图 4.18　LLMs 预测“起运港-value”类别的实体位置在 x 方向和 y 方向的概率分布

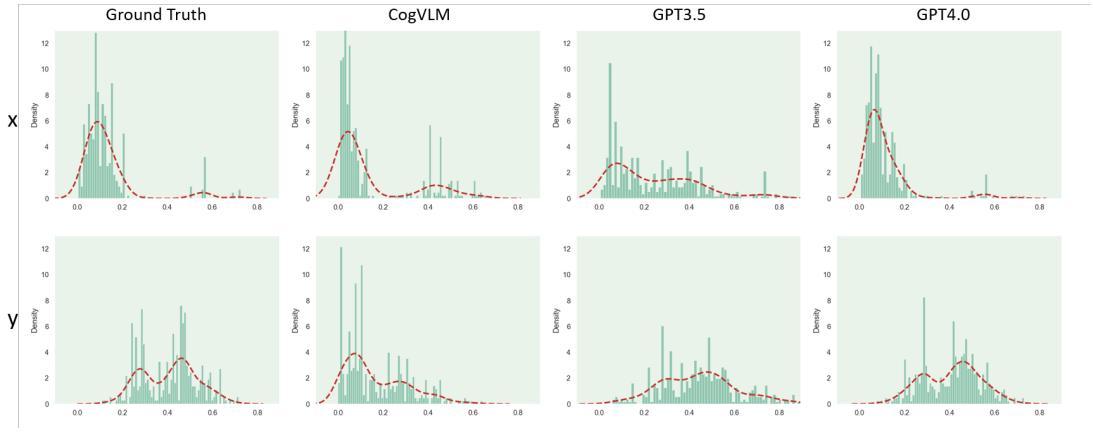


图 4.19　LLMs 预测“唛头”类别的实体位置在 x 方向和 y 方向的概率分布

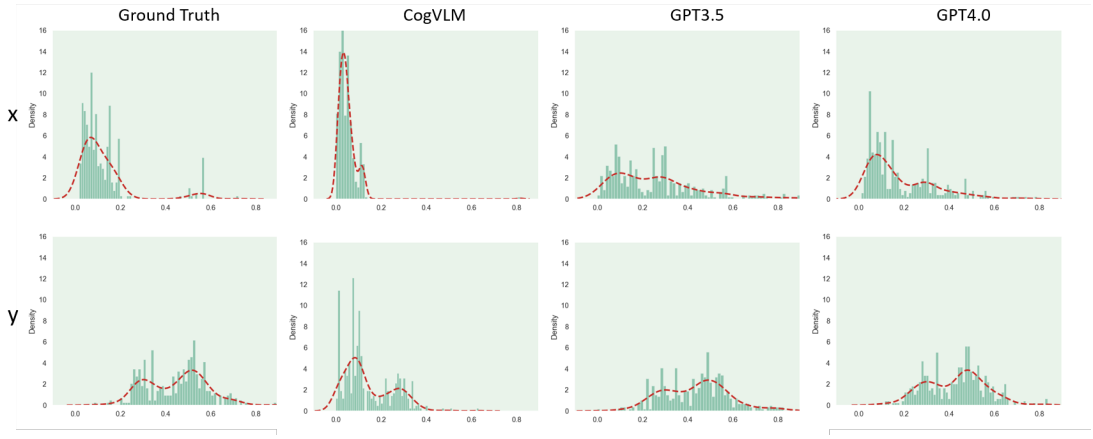


图 4.20　LLMs 预测“唛头-value”类别的实体位置在 x 方向和 y 方向的概率分布

二维散点的分布情况直观的展示了 LLMs 的布局生成能力，DLG 实验在此基础上还展示了 x 和 y 方向上的位置分布曲线，计算标准如公式 4.12 和公式 4.13 所示。图 4.15 和图 4.16 分别展示了不同 LLMs 预测“发货人”和“发货人-value”类别的实体位置在 x 和 y 方向上的概率分布曲线。图 4.17 和图 4.18 分别展示了“起运港”和“起运港-value”的分布曲线。图 4.19 和图 4.20 分别展示了“唛头”和“唛头-value”的分布曲线。从两个方向上分解二维散点分布，可以更直观的分析 LLMs 预测的实体位置在 x 方向和 y 方向的偏差程度，例如在“发货人”曲线中，可以发现在 x 方向 CogVLM 和 GPT4 的预测分布曲线与真实曲线较为相似，这说明在一定的尺度下，LLMs 对“发货人”类别在文档中具体位置的分布是与先验分布一致的。

表 4.9 LLMs 预测不同类别位置分布的 KL 一致性分数

Model	分布一致性分数（KL 散度）% ↑						
	发货人	发货人-value	起运港	起运港-value	唛头	唛头-value	Average
CogVLM-17B	37.3	46.1	50.0	50.1	20.8	47.9	42.03
ChatGPT（gpt3.5-turbo）	58.2	58.0	**89.1**	80.1	58.9	**71.6**	69.32
GPT4（gpt4-0125-preview）	**70.6**	**61.3**	85.6	**88.1**	**88.8**	67.7	**77.02**

表 4.10 LLMs 预测不同类别位置分布的 JS 一致性分数

Model	分布一致性分数（JS 散度）% ↑						
	发货人	发货人-value	起运港	起运港-value	唛头	唛头-value	Average
CogVLM-17B	86.9	87.2	85.3	84.7	83.8	87.1	85.83
ChatGPT（gpt3.5-turbo）	90.7	90.3	**97.7**	95.9	91.8	93.3	93.28
GPT4（gpt4-0125-preview）	**94.1**	**91.2**	97.0	**97.5**	**97.7**	**93.4**	**95.15**

此外，除了定量的概率分布展示，DLG 实验还定性的统计了 LLMs 预测不同类别位置分布的偏差程度，使用的评价指标是基于 KL 散度和 JS 散度的分布一致性分数。表 4.9 和表 4.10 分别展示了基于 KL 散度和 JS 散度的统计结果，分数越高表明 LLMs 预测的位置分布和真实的分布越接近。粗体表示最好的结果。从一致性分数的比较中可以发现，小型的视觉语言大模型与 ChatGPT 系列的 LLMs 在 DLG 任务上还具有一定的差距，而 ChatGPT 和 GPT4 在零样本的生成能力评估实验中展现了优秀的可塑性。在平均（Average）的实验结果中，ChatGPT 的 KL 一致性分数接近 70%，意味着模型预测的分布与真实分布大致是一致的，并

图 4.21　ChatGPT 视觉富文档布局生成的可视化案例

且由于生成任务的多样性要求，这个一致性分数基本可以具有比较良好的实际样本生成结果。此外，GPT4 在这个基础上有更好的表现。

图 4.21 展示了 ChatGPT 视觉富文档布局生成的可视化案例，图中红色和蓝色边界框分别代表生成的键值信息，边界框中的文本代表生成的文本内容，黑色字体代表生成的类别。左上为一张较为标准的 SEABILL 视觉富文档，其余三张图是 ChatGPT 根据这张图片的布局内容生成的具有相似布局的文档结构。实验要求 ChatGPT 生成包含预定义类别的边界框位置，以及边界框的文本内容。生成的布局结构的质量由专家经验总结的两条准则评价:（1）从整体看：发货人，收货人，通知人的信息一般出现在文档左上角，垂直排列；目的港，起运港一般出现在文档中间，水平排列；唛头，品名，件数，毛重等一般出现在文档最下

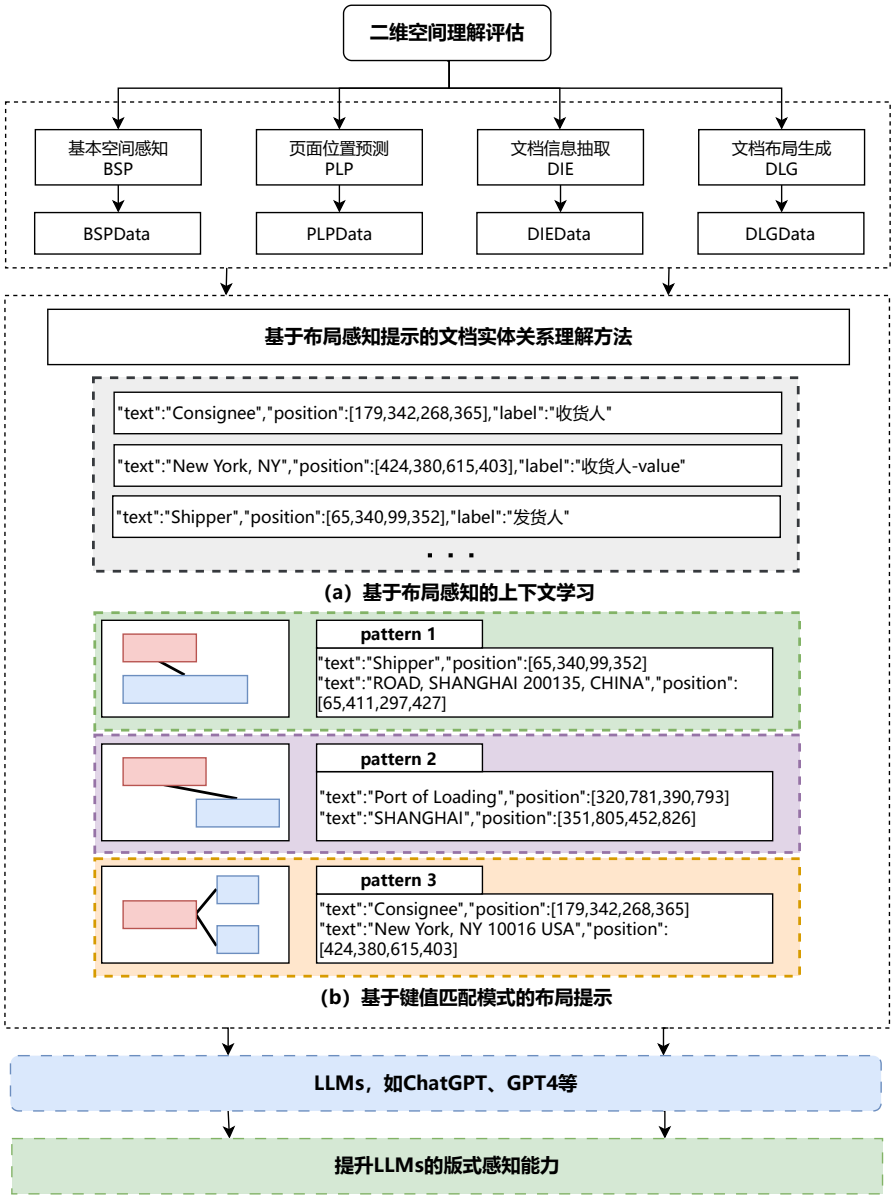


图 4.22　基于布局感知提示的文档实体关系理解方法框架

方，水平排列。(2) 从局部看；键值对之间的模式一般为正上下，斜上下或左右。从 ChatGPT 生成的布局结构来看，一些具有鲜明位置特征的类别，例如“发货人”，“发货人-value”，“起运港”，“起运港-value”，“唛头”和“唛头-value”，都基本遵循了专家经验总结的准则分布，这说明 LLMs 在 VRDs 复杂布局生成任务上具有广泛的应用前景。

4.4.4 基于布局感知提示的文档实体关系理解方法

为了解决 LLMs 布局感知能力受限的问题，本文基于 LLMs 空间位置理解评估的实验提出一种基于布局感知提示的文档实体关系理解方法，该方法主要有两个关键的部分：（1）基于布局感知的上下文学习（Layout-Aware ICL）（2）基于键值匹配模式的布局提示。如图 4.22 所示，基于布局感知提示的上下文学习提供给 LLMs 多样的学习范例，LLMs 利用强大的文本语义理解能力和空间感知能力学习并吸收这些范例，实际上，本实验提供给 LLMs 多个相似文档的 OCR 转录及对应的标注。基于键值匹配模式的布局提示认为 LLMs 学习文档实体关系应该关注特定的布局模式，例如上下排列、左右排列等，图中展示三种典型的键值排列模式，以此作为提示中约束的一部分。

表 4.11 展示了基于布局感知提示的文档实体关系理解的实验结果。与 DIE 评估任务中的零样本关系理解 RE 进行比较，可以发现提出的方法在整体上有效的提升了关系理解的效果。具体地，具有布局提示的 ChatGPT 比零样本的结果都高出超过 10 个百分点，提升幅度比 GPT4 更大，而 GPT4 的域内结果已经可以接近 90%，基本可以解决大部分视觉富文档的关系理解任务。表中的结果表明，提出的布局感知提示方法有效增强了 LLMs 的布局感知，进一步提升 LLMs 的版式感知能力。当然，基于布局感知的提示还有进一步优化的空间，设计更精妙的文本-排版跨模态提示是解决视觉富文档理解任务的关键。

表 4.11 基于布局感知提示的文档实体关系理解方法的实验结果

Setting	Model	Average			ID（域内）			OD（域外）		
		P % ↑	R % ↑	F1 % ↑	P % ↑	R % ↑	F1 % ↑	P % ↑	R % ↑	F1 % ↑
Zero-Shot	ChatGPT	53.63	53.16	53.39	76.00	47.88	58.75	52.52	51.33	51.92
	GPT4	65.76	65.52	65.64	88.58	74.46	80.91	-	-	-
Layout-Aware	ChatGPT	70.36	71.22	70.79	88.98	71.33	79.18	60.73	58.90	59.80
	GPT4	**78.71**	**78.10**	**78.40**	**94.55**	**85.49**	**89.79**	-	-	-

4.5 本章小结

本章从 LLMs 的视觉富文档理解出发，旨在探究 LLMs 的空间位置理解能力，提出了四种基于二维位置理解的零样本评估方案和数据集，从位置关系、位

置预测、关系理解和布局生成四个角度客观的评测 LLMs 的空间位置理解水平。评估实验表明 ChatGPT 系列的 LLMs 基于丰富的先验知识，能很好的理解版式布局，而较小的，具有视觉理解能力的 LLMs 在视觉富文档的理解上还有较大的局限性。此外，本章基于 LLMs 的评估实验，提出一种基于布局感知提示的文档实体关系理解方法，实验表明，该方法有效的填补了 LLMs 在版式布局理解上的偏差。本章内容的研究为 LLMs 的版式感知提供有力的研究基础和研究方向。

第五章　总结与展望

5.1　总结

随着数字化时代的快速发展，视觉富文档作为一种新型的信息传播形式，正在受到越来越多的关注和应用。相比于传统的纯文本文档，视觉富文档具有复杂的版式布局和更多的视觉元素（包括颜色、字体、边界等），这些布局和视觉信息对于文档的认知和理解起着至关重要的作用。然而，正因为众多领域的文档版式多样，采用人工的方式处理这些文档十分低效且昂贵。因此，文档智能模型的研究和自动化抽取技术的开发具有重要的应用价值。本文以国际物流运单、文档扫描件等数据为例，结合了文档中的文本内容、版式布局和视觉特征，调研并分析了国内外前沿的文档智能模型。围绕视觉富文档中视觉元素丰富、版式布局复杂等特点，本文旨在提升传统多模态视觉富文档模型的理解能力，并探索大语言模型的版式感知能力。

从跨模态提示增强的角度，本文的主要工作和贡献包括以下几个方面：

贡献一：通过视觉-排版跨模态提示增强的方法，解决多粒度视觉特征对齐、模态不均衡等问题。

为了解决基于 Transformer 架构的文档智能模型难以学习细粒度视觉特征的问题，本文提出一种新颖的基于颜色块的视觉提示方法，将不同类别的实体用不同的颜色在图片上进行填充。通过训练一个新的视觉增强网络作为”教师”模型，该模型在训练过程中由于颜色块的视觉特征，更加关注实体所对应的图像块内容和边界，从而捕捉更细粒度的视觉信息。

为了解决视觉通道的融合方式单一导致模型偏向文本、布局模态的问题，本文提出一种基于视觉-排版先验的多模态一致性学习框架，此框架在注意力机制融合的基础上，额外增加了基于颜色块的视觉信息，通过一致性学习将“教师”模型学习的细粒度特征迁移到“学生”模型，从而达到多模态均衡。大量的实验表明，提出的方法显著优于之前的多模态模型，达到目前最优的性能。

贡献二：探究大语言模型的空间感知能力，并通过文本-排版跨模态提示增强的方法，解决大语言模型空间感知能力受限的问题。

为了探究基于纯文本训练的大语言模型在视觉富文档上的空间感知能力，本文提出一套适用于大语言模型二维位置感知能力的评估方案，从四个方面评测大语言模型的位置关系理解能力，实体位置预测能力，二维键值匹配能力和版式布局生成能力。实验结果表明，提出的评估方案能客观评测大语言模型的二维空间感知能力。

为了解决二维空间的排版布局难以通过文本语义建模的问题，本文基于大语言模型的二维感知能力评估，提出一种布局感知提示的文档实体关系理解方法，将含有版式布局的文本信息作为大语言模型的上下文范例，并引入键值匹配模式的布局提示，有效提升了文档智能模型的理解能力。

5.2 展望

本文针对视觉富文档理解任务的实体语义表示和关系理解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）本文提出的 VANCL 方法是一种视觉非对称的方法，通过视觉提示增强文本、排版、视觉的多模态表示，如果利用布局非对称的一致性学习方法，应该可以进一步的提升实体的语义表示。此外 VANCL 主要在 LayoutLM 系列的模型上进行实验，没有在一些最新的主干网络上进行实验比较。

（2）本文在评估大语言模型空间感知能力时，采用的评价指标还不是最优的，比如页面位置预测 PLP 任务中使用的距离衡量，事实上，模型生成的位置包含了特别的语义，难以直接通过距离或重叠部分精确的评估。此外，本文没有使用更多的 LLMs 进行评估进行对比，这是后续需要完成的工作。

（3）本文在进行 LLMs 的实验中发现，视觉语言大模型，如 CogVLM，在输入长度较长的情况下，语言理解能力会大幅下降，导致无法给出具有意义的回答。因此，视觉编码器和文本编码器能力的平衡，这个问题值得进一步探讨。此外，视觉语言大模型虽然具有一定的多模态理解能力，但目前的大多数模型都缺乏对版式布局的理解，所以视觉语言大模型的视觉-排版跨模态提示增强的方法值得未来探索。

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] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10): 1-5.

[3] XU Y, XU Y, LV T, et al. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 2579-2591. https://aclanthology.org/2021.acl-long.201. DOI: 10.18653/v1/2021.acl-long.201.

[4] HUANG Y, LV T, CUI L, et al. LayoutLMv3: Pre-training for document ai with unified text and image masking[C/OL]//MM '22: Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2022: 4083–4091. https://doi.org/10.1145/3503161.3548112.

[5] XU Y, LI M, CUI L, et al. LayoutLM: Pre-training of text and layout for document image understanding[C/OL]//KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery, 2020: 1192–1200. https://doi.org/10.1145/3394486.3403172.

[6] HE J, WANG L, HU Y, et al. ICL-D3IE: In-context learning with diverse demonstrations updating for document information extraction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023: 19485-19494.

[7] 刘成林, 金连文, 白翔, 等. 文档智能分析与识别前沿: 回顾与展望[J]. 中国图象图形学报, 2023, 28(8): 2223-2252.

[8] GU Z, MENG C, WANG K, et al. XYLayoutLM: Towards layout-aware multimodal networks for visually-rich document understanding[C/OL]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 2022: 4573-4582. https://doi.org/10.1109/CVPR52688.2022.00454.

[9] WANG J, JIN L, DING K. LiLT: A simple yet effective language-independent layout transformer for structured document understanding[C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 7747-7757. https://aclanthology.org/2022.acl-long.534. DOI: 10.18653/v1/2022.acl-long.534.

[10] MATAS J, CHUM O, URBAN M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J/OL]. Image and Vision Computing, 2004, 22(10): 761-767. DOI: https://doi.org/10.1016/j.imavis.2004.02.006.

[11] EPSHTEIN B, OFEK E, WEXLER Y. Detecting text in natural scenes with stroke width transform[C/OL]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010: 2963-2970. DOI: 10.1109/CVPR.2010.5540041.

[12] PAN Y F, HOU X, LIU C L. A hybrid approach to detect and localize texts in natural scene images[J/OL]. IEEE Transactions on Image Processing, 2011, 20(3): 800-813. DOI: 10.1109/TIP.2010.2070803.

[13] YIN X C, YIN X, HUANG K, et al. Robust text detection in natural scene images[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 970-983. DOI: 10.1109/TPAMI.2013.182.

[14] YAO C, BAI X, LIU W, et al. Detecting texts of arbitrary orientations in natural images[C/OL]//2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012: 1083-1090. DOI: 10.1109/CVPR.2012.6247787.

[15] YE Q, DOERMANN D. Text detection and recognition in imagery: A survey[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(7): 1480-1500. DOI: 10.1109/TPAMI.2014.2366765.

[16] TIAN Z, HUANG W, HE T, et al. Detecting text in natural image with connectionist text proposal network[C]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 56-72.

[17] LIAO M, SHI B, BAI X, et al. TextBoxes: A fast text detector with a single deep neural network[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1). DOI: 10.1609/aaai.v31i1.11196.

[18] LIAO M, SHI B, BAI X. TextBoxes++: A single-shot oriented scene text detector[J/OL]. IEEE Transactions on Image Processing, 2018, 27(8): 3676-3690. DOI: 10.1109/TIP.2018.2825107.

[19] SHI B, BAI X, BELONGIE S. Detecting oriented text in natural images by linking segments[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[20] LIU Y, CHEN H, SHEN C, et al. ABCNet: Real-time scene text spotting with adaptive bezier-curve network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.

[21] ZHU Y, CHEN J, LIANG L, et al. Fourier contour embedding for arbitrary-shaped text detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 3123-3131.

[22] ZHANG Z, ZHANG C, SHEN W, et al. Multi-oriented text detection with fully convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[23] DENG D, LIU H, LI X, et al. PixelLink: Detecting scene text via instance segmentation [J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1). DOI: 10.1609/aaai.v32i1.12269.

[24] WANG W, XIE E, LI X, et al. Shape robust text detection with progressive scale expansion network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[25] XU Y, WANG Y, ZHOU W, et al. TextField: Learning a deep direction field for irregular scene text detection[J/OL]. IEEE Transactions on Image Processing, 2019, 28(11): 5566-5579. DOI: 10.1109/TIP.2019.2900589.

[26] BAEK Y, LEE B, HAN D, et al. Character region awareness for text detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[27] ZHOU X, YAO C, WEN H, et al. EAST: An efficient and accurate scene text detector[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[28] HE W, ZHANG X Y, YIN F, et al. Multi-oriented and multi-lingual scene text detection with direct regression[J/OL]. IEEE Transactions on Image Processing, 2018, 27(11): 5406-5419. DOI: 10.1109/TIP.2018.2855399.

[29] ZHANG S X, ZHU X, YANG C, et al. Adaptive boundary proposal network for arbitrary shape text detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 1305-1314.

[30] TANG J, ZHANG W, LIU H, et al. Few could be better than all: Feature sampling and grouping for scene text detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 4563-4572.

[31] SONG S, WAN J, YANG Z, et al. Vision-language pre-training for boosting scene text detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 15681-15691.

[32] BISSACCO A, CUMMINS M, NETZER Y, et al. PhotoOCR: Reading text in uncontrolled conditions[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2013.

[33] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2298-2304. DOI: 10.1109/TPAMI.2016.2646371.

[34] YIN F, WU Y C, ZHANG X Y, et al. Scene text recognition with sliding convolutional character models[A]. 2017. arXiv: 1709.01727.

[35] SHI B, WANG X, LYU P, et al. Robust scene text recognition with automatic rectification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[36] LI H, WANG P, SHEN C, et al. Show, attend and read: A simple and strong baseline for irregular text recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 8610-8617.

[37] CHENG Z, BAI F, XU Y, et al. Focusing attention: Towards accurate text recognition in natural images[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017.

[38] WANG T, ZHU Y, JIN L, et al. Decoupled attention network for text recognition[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07): 12216-12224. DOI: 10.1609/aaai.v34i07.6903.

[39] SHI B, YANG M, WANG X, et al. ASTER: An attentional scene text recognizer with flexible rectification[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2035-2048. DOI: 10.1109/TPAMI.2018.2848939.

[40] LUO C, JIN L, SUN Z. MORAN: A multi-object rectified attention network for scene text recognition[J/OL]. Pattern Recognition, 2019, 90: 109-118. DOI: https://doi.org/10.1016/j.patcog.2019.01.020.

[41] FANG S, XIE H, WANG Y, et al. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 7098-7107.

[42] BAUTISTA D, ATIENZA R. Scene text recognition with permuted autoregressive sequence models[C]//AVIDAN S, BROSTOW G, CISSÉ M, et al. Computer Vision – ECCV 2022. Cham: Springer Nature Switzerland, 2022: 178-196.

[43] DU Y, CHEN Z, JIA C, et al. Svtr: Scene text recognition with a single visual model[A]. 2022. arXiv: 2205.00159.

[44] LI H, WANG P, SHEN C. Towards end-to-end text spotting with convolutional recurrent neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017.

[45] LIU X, LIANG D, YAN S, et al. FOTS: Fast oriented text spotting with a unified network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[46] FENG W, HE W, YIN F, et al. TextDragon: An end-to-end framework for arbitrary shaped text spotting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019.

[47] LIAO M, PANG G, HUANG J, et al. Mask textSpotter v3: Segmentation proposal network for robust scene text spotting[C]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 706-722.

[48] LYU P, LIAO M, YAO C, et al. Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[49] HUANG M, LIU Y, PENG Z, et al. SwinTextSpotter: Scene text spotting via better synergy between text detection and text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 4593-4603.

[50] XING L, TIAN Z, HUANG W, et al. Convolutional character networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019.

[51] QIAO L, CHEN Y, CHENG Z, et al. MANGO: A mask attention guided one-stage scene text spotter[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2467-2476. DOI: 10.1609/aaai.v35i3.16348.

[52] ZHANG X, SU Y, TRIPATHI S, et al. Text spotting transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 9519-9528.

[53] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 213-229.

[54] PENG D, WANG X, LIU Y, et al. SPTS: Single-point text spotting[C/OL]//MM '22: Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2022: 4272–4281. https://doi.org/10.1145/3503161.3547942.

[55] WONG K Y, CASEY R G, WAHL F M. Document analysis system[J/OL]. IBM Journal of Research and Development, 1982, 26(6): 647-656. DOI: 10.1147/rd.266.0647.

[56] KISE K, SATO A, IWATA M. Segmentation of page images using the area voronoi diagram [J/OL]. Computer Vision and Image Understanding, 1998, 70(3): 370-382. DOI: https://doi.org/10.1006/cviu.1998.0684.

[57] GAO L, YI X, JIANG Z, et al. ICDAR2017 competition on page object detection[C/OL]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): Vol. 01. 2017: 1417-1422. DOI: 10.1109/ICDAR.2017.231.

[58] REN S, HE K, GIRSHICK R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C/OL]//CORTES C, LAWRENCE N D, LEE D D, et al. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. 2015: 91-99. https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

[59] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[60] YANG X, YUMER E, ASENTE P, et al. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[61] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks [A]. 2017. arXiv: 1609.02907.

[62] VELIčKOVIć P, CUCURULL G, CASANOVA A, et al. Graph attention networks[A]. 2018. arXiv: 1710.10903.

[63] WANG Z, XU Y, CUI L, et al. LayoutReader: Pre-training of text and layout for reading order detection[A]. 2021. arXiv: 2108.11591.

[64] SIDDIQUI S A, FATEH I A, RIZVI S T R, et al. DeepTabStR: Deep learning based table structure recognition[C/OL]//2019 International Conference on Document Analysis and Recognition (ICDAR). 2019: 1403-1409. DOI: 10.1109/ICDAR.2019.00226.

[65] LI Y, HUANG Y, ZHU Z, et al. Rethinking table structure recognition using sequence labeling methods[C]//LLADÓS J, LOPRESTI D, UCHIDA S. Document Analysis and Recognition – ICDAR 2021. Cham: Springer International Publishing, 2021: 541-553.

[66] ZHONG X, SHAFIEIBAVANI E, JIMENO YEPES A. Image-based table recognition: Data, model, and evaluation[C]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 564-580.

[67] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[C]//LEEN T, DIETTERICH T, TRESP V. Advances in Neural Information Processing Systems: Vol. 13. MIT Press, 2000.

[68] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12: 2493-2537.

[69] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//BURGES C, BOTTOU L, WELLING M, et al. Advances in Neural Information Processing Systems: Vol. 26. Curran Associates, Inc., 2013.

[70] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[A]. 2013. arXiv: 1301.3781.

[71] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[72] DAI A M, LE Q V. Semi-supervised sequence learning[C]//CORTES C, LAWRENCE N, LEE D, et al. Advances in Neural Information Processing Systems: Vol. 28. Curran Associates, Inc., 2015.

[73] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C/OL]//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.

[74] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[M]. OpenAI, 2018.

[75] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]// LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems: Vol. 33. Curran Associates, Inc., 2020: 1877-1901.

[76] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//KOYEJO S, MOHAMED S, AGARWAL A, et al. Advances in Neural Information Processing Systems: Vol. 35. Curran Associates, Inc., 2022: 27730-27744.

[77] DONG Q, LI L, DAI D, et al. A survey on in-context learning[A]. 2022.

[78] 周飞燕, 金林鹏, 董军, 等. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.

[79] DING L, LI H, HU C, et al. AlexNet feature extraction and multi-kernel learning for objectoriented classification[J/OL]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2018, XLII-3: 277-281. DOI: 10.5194/isprs-archives-XLII-3-277-2018.

[80] 赵立新, 邢润哲, 白银光, 等. 深度学习在目标检测的研究综述[J]. 科学技术与工程, 2021, 21(30): 12787-12795.

[81] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[A]. 2014.

[82] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[83] TAO S, LI Y, HUANG Y, et al. Face detection algorithm based on deep residual network [J/OL]. Journal of Physics: Conference Series, 2021, 1802(3): 032142. DOI: 10.1088/1742-6596/1802/3/032142.

[84] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//GUYON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.

[85] DENGEL A R, KLEIN B. smartFIX: A requirements-driven system for document analysis and understanding[C]//LOPRESTI D, HU J, KASHI R. Document Analysis Systems V. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002: 433-444.

[86] SCHUSTER D, MUTHMANN K, ESSER D, et al. Intellix – end-user trained information extraction for document archiving[C/OL]//2013 12th International Conference on Document Analysis and Recognition. 2013: 101-105. DOI: 10.1109/ICDAR.2013.28.

[87] 张言, 李强, 申化文, 等. 以文字为中心的图像理解技术综述[J]. 中国图象图形学报, 2023, 28(8): 2253-2275.

[88] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: Attentive language models beyond a fixed-length context[A]. 2019. arXiv: 1901.02860.

[89] KATTI A R, REISSWIG C, GUDER C, et al. Chargrid: Towards understanding 2D documents [C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 4459-4469. https://aclanthology.org/D18-1476. DOI: 10.18653/v1/D18-1476.

[90] KERROUMI M, SAYEM O, SHABOU A. VisualWordGrid: Information extraction from scanned documents using a multimodal approach[C/OL]//BARNEY SMITH E H, PAL U. Document Analysis and Recognition – ICDAR 2021 Workshops. Cham: Springer International Publishing, 2021: 389-402. DOI: 10.1007/978-3-030-86159-9_28.

[91] LIN W, GAO Q, SUN L, et al. ViBERTgrid: A jointly trained multi-modal 2d document representation for key information extraction from documents[C/OL]//LLADÓS J, LOPRESTI D, UCHIDA S. Document Analysis and Recognition – ICDAR 2021. Cham: Springer International Publishing, 2021: 548-563. https://arxiv.org/pdf/2105.11672.pdf.

[92] LIU X, GAO F, ZHANG Q, et al. Graph convolution for multimodal information extraction from visually rich documents[C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 32-39. https://aclanthology.org/N19-2005. DOI: 10.18653/v1/N19-2005.

[93] YU W, LU N, QI X, et al. PICK: Processing key information extraction from documents using improved graph learning-convolutional networks[C/OL]//2020 25th International Conference on Pattern Recognition (ICPR). 2021: 4363-4370. DOI: 10.1109/ICPR48806.2021.9412927.

[94] QIAN Y, SANTUS E, JIN Z, et al. GraphIE: A graph-based framework for information extraction[C/OL]//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: 751-761. https://aclanthology.org/N19-1082. DOI: 10.18653/v1/N19-1082.

[95] HUANG Z, CHEN K, HE J, et al. ICDAR2019 competition on scanned receipt ocr and information extraction[C/OL]//2019 International Conference on Document Analysis and Recognition (ICDAR). 2019: 1516-1520. DOI: 10.1109/ICDAR.2019.00244.

[96] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C/OL]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016: 260-270. https://aclanthology.org/N16-1030. DOI: 10.18653/v1/N16-1030.

[97] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized bert pretraining approach [A]. 2019. arXiv: 1907.11692.

[98] PALM R B, WINTHER O, LAWS F. CloudScan - a configuration-free invoice analysis system using recurrent neural networks[J/OL]. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, 01: 406-413. http://arxiv.org/abs/1708.07403.

[99] SAGE C, AUSSEM A, ELGHAZEL H, et al. Recurrent neural network approach for table field extraction in business documents[C/OL]//2019 International Conference on Document Analysis and Recognition (ICDAR). 2019: 1308-1313. DOI: 10.1109/ICDAR.2019.00211.

[100] HWANG W, KIM S, SEO M, et al. Post-OCR parsing: building simple and robust parser via {bio} tagging[C/OL]//Workshop on Document Intelligence at NeurIPS 2019. 2019. https://openreview.net/forum?id=SJgjf695UB.

[101] DENK T I, REISSWIG C. BERTgrid: Contextualized embedding for 2d document representation and understanding[C/OL]//Workshop on Document Intelligence at NeurIPS 2019. 2019. https://openreview.net/forum?id=H1gsGaq9US.

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

[103] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [A]. 2017. arXiv: 1707.06347.

[104] LIU X, JI K, FU Y, et al. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 61-68. https://aclanthology.org/2022.acl-short.8. DOI: 10.18653/v1/2022.acl-short.8.

[105] HU E J, YELONG SHEN, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[C/OL]//International Conference on Learning Representations. 2022. https://openreview.net/forum?id=nZeVKeeFYf9.

[106] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[A]. 2023. arXiv: 2303.18223.

[107] HO N, SCHMID L, YUN S Y. Large language models are reasoning teachers[A]. 2022.

[108] WEI J, WANG X, SCHUURMANS D, et al. Chain of thought prompting elicits reasoning in large language models[A]. 2022.

[109] 罗文. 大语言模型评测综述[J]. 中文信息学报, 2024, 38(1): 1-23.

[110] LIANG P, BOMMASANI R, LEE T, et al. Holistic evaluation of language models[A]. 2023. arXiv: 2211.09110.

[111] LIU H, LI C, WU Q, et al. Visual instruction tuning[A]. 2023. arXiv: 2304.08485.

[112] WANG W, LV Q, YU W, et al. CogVLM: Visual expert for large language models[EB/OL]. 2024. https://openreview.net/forum?id=c72vop46KY.

[113] GARNCAREK Ł, POWALSKI R, STANISŁAWEK T, et al. LAMBERT: Layout-aware language modeling for information extraction[C/OL]//LLADÓS J, LOPRESTI D, UCHIDA S. Document Analysis and Recognition – ICDAR 2021. Cham: Springer International Publishing, 2021: 532-547. DOI: 10.1007/978-3-030-86549-8_34.

[114] GU J, KUEN J, MORARIU V I, et al. Unidoc: Unified pretraining framework for document understanding[J/OL]. Advances in Neural Information Processing Systems, 2021, 34: 39-50. https://proceedings.neurips.cc/paper/2021/hash/0084ae4bc24c0795d1e6a4f58444d39b-Abstract.html.

[115] LI P, GU J, KUEN J, et al. Selfdoc: Self-supervised document representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5652-5660.

[116] LI Y, QIAN Y, YU Y, et al. StrucTexT: Structured text understanding with multi-modal transformers[C/OL]//MM '21: Proceedings of the 29th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2021: 1912–1920. https://doi.org/10.1145/3474085.3475345.

[117] APPALARAJU S, JASANI B, KOTA B U, et al. Docformer: End-to-end transformer for document understanding[C/OL]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 993-1003. https://arxiv.org/abs/2106.11539.

[118] LEE C Y, LI C L, DOZAT T, et al. FormNet: Structural encoding beyond sequential modeling in form document information extraction[C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 3735-3754. https://aclanthology.org/2022.acl-long.260. DOI: 10.18653/v1/2022.acl-long.260.

[119] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C/OL]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778. http://arxiv.org/abs/1512.03385.

[120] COONEY C, HEYBURN R, MADIGAN L, et al. Unimodal and multimodal representation training for relation extraction[C/OL]//LONGO L, O'REILLY R. Artificial Intelligence and Cognitive Science. Cham: Springer Nature Switzerland, 2023: 450-461. https://link.springer.com/chapter/10.1007/978-3-031-26438-2_35.

[121] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 8440-8451. https://aclanthology.org/2020.acl-main.747. DOI: 10.18653/v1/2020.acl-main.747.

[122] CHI Z, DONG L, WEI F, et al. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training[C/OL]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 3576-3588. https://aclanthology.org/2021.naacl-main.280. DOI: 10.18653/v1/2021.naacl-main.280.

[123] TANG G, XIE L, JIN L, et al. MatchVIE: Exploiting match relevancy between entities for visual information extraction[C/OL]//ZHOU Z H. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conferences on Artificial Intelligence Organization, 2021: 1039-1045. https://doi.org/10.24963/ijcai.2021/144.

[124] POWALSKI R, BORCHMANN Ł, JURKIEWICZ D, et al. Going full-tilt boogie on document understanding with text-image-layout transformer[C/OL]//Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16. Springer, 2021: 732-747. https://link.springer.com/chapter/10.1007/978-3-030-86331-9_47.

[125] WANG Z, GU J, TENSMEYER C, et al. MGDoc: Pre-training with multi-granular hierarchy for document image understanding[A/OL]. 2022. https://arxiv.org/abs/2211.14958.

[126] YU Y, LI Y, ZHANG C, et al. StrucTexTv2: Masked visual-textual prediction for document image pre-training[J/OL]. CoRR, 2023, abs/2303.00289. https://doi.org/10.48550/arXiv.2303.00289.

[127] MIYATO T, MAEDA S I, KOYAMA M, et al. Virtual adversarial training: A regularization method for supervised and semi-supervised learning[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1979-1993. DOI: 10.1109/TPAMI.2018.2858821.

[128] XIE Q, DAI Z, HOVY E, et al. Unsupervised data augmentation for consistency training [C/OL]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems: Vol. 33. Curran Associates, Inc., 2020: 6256-6268. https://proceedings.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf.

[129] LOWELL D, HOWARD B, LIPTON Z C, et al. Unsupervised data augmentation with naive augmentation and without unlabeled data[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021a: 4992-5001. https://aclanthology.org/2021.emnlp-main.408. DOI: 10.18653/v1/2021.emnlp-main.408.

[130] CHEN J, SHEN D, CHEN W, et al. HiddenCut: Simple data augmentation for natural language understanding with better generalizability[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 4380-4390. https://aclanthology.org/2021.acl-long.338. DOI: 10.18653/v1/2021.acl-long.338.

[131] LOWELL D, HOWARD B, LIPTON Z C, et al. Unsupervised data augmentation with naive augmentation and without unlabeled data[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021b: 4992-5001. https://aclanthology.org/2021.emnlp-main.408. DOI: 10.18653/v1/2021.emnlp-main.408.

[132] BATRA T, PARIKH D. Cooperative learning with visual attributes[J/OL]. CoRR, 2017, abs/1705.05512. http://arxiv.org/abs/1705.05512.

[133] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 4320-4328. http://arxiv.org/abs/1706.00384.

[134] HINTON G E, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J/OL]. CoRR, 2015, abs/1503.02531. http://arxiv.org/abs/1503.02531.

[135] REN Y, XIONG D. CogAlign: Learning to align textual neural representations to cognitive language processing signals[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 3758-3769. https://aclanthology.org/2021.acl-long.291. DOI: 10.18653/v1/2021.acl-long.291.

[136] LIANG X, WU L, LI J, et al. R-drop: Regularized dropout for neural networks[C/OL]// RANZATO M, BEYGELZIMER A, DAUPHIN Y, et al. Advances in Neural Information Processing Systems: Vol. 34. Curran Associates, Inc., 2021: 10890-10905. https://proceedings.neurips.cc/paper/2021/file/5a66b9200f29ac3fa0ae244cc2a51b39-Paper.pdf.

[137] CHEN X, YUAN Y, ZENG G, et al. Semi-supervised semantic segmentation with cross pseudo supervision[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 2613-2622. DOI: 10.1109/CVPR46437.2021.00264.

[138] YANG P, CONG X, SUN Z, et al. Enhanced language representation with label knowledge for span extraction[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 4623-4635. https://aclanthology.org/2021.emnlp-main.379. DOI: 10.18653/v1/2021.emnlp-main.379.

[139] JAUME G, KEMAL EKENEL H, THIRAN J P. FUNSD: A dataset for form understanding in noisy scanned documents[C/OL]//2019 International Conference on Document Analysis and Recognition Workshops (ICDARW): Vol. 2. 2019: 1-6. DOI: 10.1109/ICDARW.2019.10029.

[140] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized bert pretraining approach [EB/OL]. 2019. https://arxiv.org/abs/1907.11692. DOI: 10.48550/ARXIV.1907.11692.

[141] BAO H, DONG L, WEI F, et al. UniLMv2: Pseudo-masked language models for unified language model pre-training[C/OL]//III H D, SINGH A. Proceedings of Machine Learning Research: Vol. 119 Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020: 642-652. https://proceedings.mlr.press/v119/bao20a.html.

[142] HONG T, KIM D, JI M, et al. BROS: A pre-trained language model for understanding texts in document[EB/OL]. 2021. https://openreview.net/forum?id=punMXQEsPr0.

[143] LIN J. Divergence measures based on the shannon entropy[J/OL]. IEEE Transactions on Information Theory, 1991, 37(1): 145-151. DOI: 10.1109/18.61115.

[144] KULLBACK S, LEIBLER R A. On information and sufficiency[J/OL]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86[2022-10-17]. http://www.jstor.org/stable/2236703.

[145] GEGENFURTNER K R. Cortical mechanisms of colour vision[J/OL]. Nature Reviews Neuroscience, 2003, 4(7): 563-572. https://doi.org/10.1038/nrn1138.

[146] ELLIOT A J, MAIER M A. Color psychology: effects of perceiving color on psychological functioning in humans.[J/OL]. Annual review of psychology, 2014, 65: 95-120. https://deweycolorsystem.com/wp-content/uploads/2020/06/Credentials-Color-Psychology.pdf.

[147] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J/OL]. Proceedings of the IEEE, 1998, 86(11): 2278-2324. DOI: 10.1109/5.726791.

[148] BANG Y, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity[A]. 2023.

[149] DONG Q, LI L, DAI D, et al. A survey for in-context learning[A]. 2022.

[150] SONG M, JIANG H, SHI S, et al. Is chatgpt a good keyphrase generator? a preliminary study [A]. 2023.

[151] MIN S, LEWIS M, ZETTLEMOYER L, et al. MetaICL: Learning to learn in context[C/OL]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, 2022: 2791-2809. https://aclanthology.org/2022.naacl-main.201. DOI: 10.18653/v1/2022.naacl-main.201.

作者在攻读硕士学位期间发表的论文与研究成果

一、发表论文

[1] H. Wang, **X. Chen**, R. Wang*, and C. Chu. 2023. Vision-Enhanced Semantic Entity Recognition in Document Images via Visually-Asymmetric Consistency Learning. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing **(EMNLP 2023)**, pages 15718–15731, Singapore. Association for Computational Linguistics.（第二作者, 导师一作, CCF-B, *Accepted*）

[2] **X. Chen**, et al.. Exploring Spatial Understanding Capability in Large Language Models: Proficiency in Layout Generation sans Visual Perception. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing **(EMNLP 2024)**.（第一作者, CCF-B, *submitted*）

[3] 李唐, 张俊伟, **陈夏华**, 王昊 *. 面向国际物流领域的视觉富文档数据集构建及信息抽取方法研究, 第十六届全国知识图谱与语义计算大会 (CCKS), 2022, Poster.（第三作者）

二、参与项目

1. 国家自然科学基金青年项目“基于人机协同拟态学习的富文档隐性模式与知识获取研究”(项目编号：62306173), 参与, 2024-01-01 至 2026-12-31。

2. 上海市科技创新行动扬帆计划项目，“药物警戒中跨语言多知识驱动的不良事件抽取方法研究”(项目编号：21YF1413900), 参与, 2021-05 至 2024-04。

3. 横向研发项目-企业委托，“基于多模态预训练模型的国际物流表单结构化抽取方法研究”, 参与, 2021-10 至 2022-10。

致　谢

在我即将完成我的硕士学位之际，我想借此机会向所有在我硕士学习生涯中给予我支持和帮助的人们致以最诚挚的感谢和深深的敬意。

首先，我要感谢王老师和朱老师。在整个硕士学习期间，两位老师不仅在学术上给予了我悉心的指导和启发，还在生活上给予了我无私的关怀和支持。老师们的深厚学识、严谨治学的态度以及对科研的热情不仅激励着我不断进步，更让我受益匪浅。

我要感谢实验室的各位老师和同学们。正是在这个团结友爱的大家庭中，我不仅学到了专业知识，更锻炼了自己的实践能力和团队合作精神。特别要感谢我的两位师弟，他们在我论文研究的过程中给予了我许多的建议和帮助。

我还要感谢家人。在我漫长的求学之路上，是他们始终给予我无限的支持和理解，是他们的支持让我能够坚持不懈地追求自己的梦想。我要特别感谢我的父母，是他们的辛勤付出和教导让我成为了今天的我。

此外，我还要感谢所有帮助过我的老师、同学和朋友们，是你们的鼓励和支持让我能够克服各种困难，坚定地走向成功。

最后，我要感谢祖国和时代。正是在祖国政策的支持下，我才有了接受高等教育的机会；正是在这个蓬勃发展的时代，我才有了施展才华的舞台。我将铭记时代赋予我的使命，为国家的繁荣进步贡献自己的力量。

再次感谢所有曾经帮助过我的人们，是你们的支持和帮助让我能够顺利完成我的硕士学业。我将永远怀着感激之情，铭记于心。

谨以此文，向您们致以最诚挚的感谢！

陈夏华

上海大学计算机工程与科学学院

2024 年 05 月 29 日