arXiv:2512.18122v1 [cs.MM] 19 Dec 2025

通过辅助生成加速端到端 PDF 到 Markdown 的转换

段长旭[0000−0003−0547−0901]

达姆施塔特工业大学
duan@linglit.tu-darmstadt.de

摘要。 将 PDF 等机器无法读取的格式中的数据转换为 Markdown，有望提升科学研究的可访问性。现有的端到端解码器 Transformer 模型能够将 PDF 截图转换为 Markdown，相较于基于流水线的方法更加灵活。然而，从头开始逐个 token 解码效率低下，尤其是在 PDF 中的密集文本可以直接复制时。为应对这一挑战，本文对提示查找解码（PLD）进行了修改，可直接从 PDF 文件中提取候选序列，利用 PDF 与其 Markdown 对应内容之间较高的 n 元语法重叠。本文提出了一种新方法——复制查找解码（CLD），以改进 PLD 的候选生成机制。实验表明，CLD 在保持原有质量的同时，可将转换过程加速至多 1.70× 。本文的代码库已在 GitHub 上开源 [1]。

关键词： PDF 转 Markdown · 辅助生成 · 文档布局分析

1 引言

科学研究成果几乎全部以 PDF 格式发表。PDF 是一种非结构化文本格式，其内容不易被机器读取。PDF 格式的科学文档包含数学公式、图形、页眉和表格等复杂元素，文本排版也十分紧密。视觉语言模型（VLM）能够理解以文档截图形式呈现的文档，但其理解能力尚不够可靠 [9]。

由于语法简单且格式结构化，Markdown 在机器可读性方面是一种理想格式，能够保留上下文并增强可理解性。Markdown 支持粗体、斜体和超链接等富文本元素，可丰富数据并为模型生成准确输出提供更详尽的上下文。此外，Markdown 社区扩展了其语法，使其能够支持表格以及高级数学公式和化学公式 [23]。因此，将 PDF 文件转换为 Markdown 值得投入精力。

一种广泛使用的 PDF 转 Markdown 方法采用基于流水线的方法，如 Marker [37] 所示。该流程首先进行

[1] https://github.com/Fireblossom/CopyLookup

文档版面分析，将每个 PDF 页面划分为多个区块，并识别每个区块的类型。随后，Marker 确定这些区块的阅读顺序。根据区块类型，系统会调用不同模块，将公式和表格等特定元素转换为 Markdown。最后，流程会经过清理阶段以纠正错误。然而，基于流水线的方法在处理复杂版面的文档时可能会遇到困难，因为某个步骤中的错误可能会传播并在后续步骤中累积，从而影响整体性能，如图1所示。

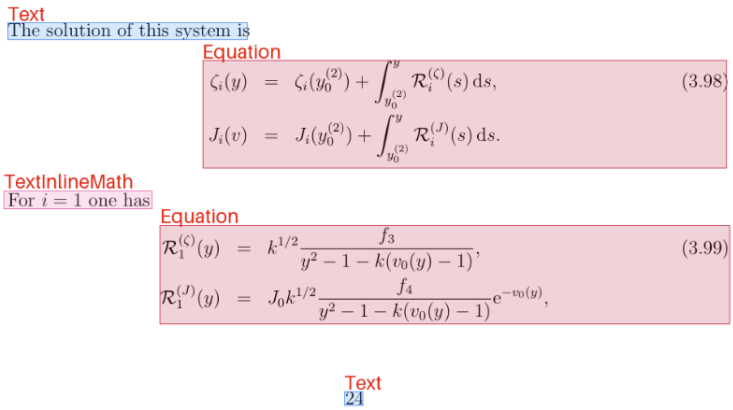


图1：Marker 对复杂版式科学文档进行版面分析的示例。Marker 能够正确检测数学公式和普通文本行。然而，行内数学表达式及其附近的普通文本行被错误地合并；页码也被识别为普通文本。文档版面分析中的错误会逐步累积，并传播到流水线的后续步骤。

基于流水线的方法之外，另一种方案是端到端解码器 Transformer 模型，例如 Nougat [4]。该模型是一种小型视觉语言模型，可读取 PDF 页面的屏幕截图，并直接将内容提取为 Markdown，自动保留正确的阅读顺序，并将数学公式转换为 LaTeX 代码。然而，Nougat 在处理速度方面存在局限，因为它从头开始逐个生成文本 token，这可能会显著拖慢文档转换过程。

辅助生成 [14,16]可以在不降低生成质量的情况下加速 Transformer 解码器模型，因为所有候选内容都会由 Transformer模型验证。提示查找解码（PLD） [31]在输入（提示）与大语言模型的输出之间存在大量 n 元语法重叠时，尤其有优势。这种方法在摘要和上下文问答等任务中可提速至 2.4× 倍。鉴于PDF文件经常——

常包含适合直接复制的密集文本，因此实施 PLD 可以简化转换流程并降低计算成本。

本文的贡献包括：

1. 改进原始 PLD（mPLD），使其能够从 PDF 文件中提取候选序列。mPLD 即插即用，无需微调，并且兼容任何端到端 PDF 转 Markdown 转换模型。2. 提出复制查找解码（CLD），以增强 mPLD 中的候选生成。3. 发布数据集制作脚本以及实验所用文档的 arXiv 编号，以增强可复现性并遵守版权规定。

2 相关工作

2.1 基于流水线的 PDF 到 Markdown 转换

GROBID [1]是一种机器学习库，可用于提取、解析和重构文档（包括 PDF），并将其编码为结构化 XML 文档，之后可转换为 Markdown。然而，它不够灵活，因为会将公式和表格转换为图像，从而妨碍后续的访问。DocTR [25] 和 DocBed [46] 首先识别文档版面，然后提取文本内容。Marker、MinerU [6] 和 Docling [36] 通过文本版面分析、阅读顺序检测、公式或表格提取，最终将这些内容整合为 Markdown，实现 PDF 到 Markdown 转换。一些类似 LlamaParse[2] 和 Mistral OCR [3] 的闭源在线服务也提供 PDF 到 Markdown 转换，其流程与文档中所述的 Marker 类似。

2.2 端到端 PDF 到 Markdown 转换

端到端 PDF 到 Markdown 转换模型基于视觉语言模型（VLM）和文档理解（DU）Transformer 模型构建。Donut[15] 是一种 DU 模型，由视觉编码器和语言模型解码器组成，不会直接从文档中提取文本。Donut 的预训练目标是复述文档截图中的文本。Nougat [4] 采用来源于 arXiv 的数据集，实现了学术文档的截图转 Markdown。 μgat [30] 扩展了 Nougat 的输入，使其能够处理跨越单页限制的元素。LOCR [35]通过额外的位置提示，解决了 Nougat 的幻觉和重复问题。Kosmos-2.5 [22] 通过更大的模型规模，实现了通用的截图转 Markdown。本文介绍的方法旨在加速 Nougat 和 Kosmos-2.5。

通用视觉语言模型也用于实现截图到 Markdown 的转换。LlamaOCR [24], Zerox OCR [2]和 Parallex [34] 都会将截图

[2] https://cloud.llamaindex.ai/parse[3] https://mistral.ai/news/mistral-ocr

输入给 PDF 的开源和闭源视觉语言模型，并通过传入恰当的提示词，引导模型返回所需输出。在本文中，我们复现了他们的方法，并使用两种现成的开源模型 Llama3.2 [3] 和 Qwen2-VL [38]作为骨干模型，开展加速实验。

2.3 文档版面分析

近年来，得益于大规模文档布局数据集 [13,18,28,45]的出现，文档版面分析（DLA）模型的能力日益增强。计算机视觉模型能够从文档截图中提取版面 [17,39,42]。语言模型也已应用于版面识别。LayoutLM [41] 及其变体 VILA [32]是 Transformer 编码器模型，可根据文本及其二维坐标分析文档布局。LayoutLMv2 和 v3 [12, 40] 还将视觉特征融入 Transformer 编码器。

在本研究中，我们的 CLD 可复制文本识别模块由 ERNIE Layout [27]微调而来。

2.4 辅助生成

为缓解大语言模型（LLM）自回归解码所导致的高推理延迟，辅助生成 [16]已成为一种用于大语言模型推理的新型解码范式。在每个解码步骤中，该方法首先利用一个较小的大语言模型高效地草拟若干后续词元，然后并行验证这些词元。PLD [31] 和 LLMA [43] 将末尾的几个词元与输入提示中较早位置的内容进行匹配，并选取相应的文本片段作为候选项。LookaheadDecoding [10] 在输入提示末尾引入多个特殊词元，以实现并行草拟，并将草稿转换为 n 元语法候选项。Lookahead [44] 采用多分支草拟策略。由于草拟阶段的时间开销较低，这些方法能够持续提升目标大语言模型的推理速度。

近期的一些研究 [5, 19, 20] 将大语言模型的隐藏状态用作其草拟模型的输入。然而，这些方法需要额外的参数，并且需要针对每个骨干模型进行训练。在这项工作中，我们改进了 PLD，以展示我们方法的有效性，因为 PLD 是最容易实现的方法，并且已集成到 Hugging Face库中。

3 方法

首先，PLD 会在 Transformer模型的输入（提示）中查找与当前预测匹配的 n 元语法，并直接从提示中复制若干词元作为候选词元。随后，大语言模型会同时验证所有候选词元，从左到右依次判断它们是否可能是序列中的下一个词元，直到遇到第一个不合理的候选词元。然后，Transformer模型会接受最后一段被判定为

合理的序列。改进的 PLD（mPLD）只需进行一项简单调整：将候选序列的来源从提示改为 PDF 页面上的全部文本。

然而，这种修改带来了两个重大问题。首先，PDF 中并非所有文本都能轻易复制，这可能会限制提取效果。其次，默认情况下，mPLD 会从页面上第一个匹配的 n 元语法位置提取候选序列，因此有可能无法正确匹配文档后部的候选项。

为了解决这些问题，我们设计了复制查找解码（CLD），通过可复制文本识别和候选生成来过滤可能的 PDF 噪声，并将候选项与页面后部的内容进行匹配。

3.1 可复制文本识别

受 DLA 相关研究的启发，我们认为文本是否可复制与其版面信息相关。具体而言，我们提出：

1. 段落中的纯文本应当可复制。2. 数学公式、表格和图形等页面元素**不**应可复制。3. 页眉、页脚和页码等不传达相关内容的元素**不**应可复制。

遵循 VILA [32] ，，我们假设文本是否可复制在文本片段级别上具有一致性。我们使用 PyMuPDF [29]从 PDF 中提取文本片段级文本和边界框。PyMuPDF 通常能保留正确的阅读顺序，但脚注和图注等浮动元素的情况除外。我们使用 LoRA [11] 和一个非官方实现 [26] 对 ERNIE Layout 模型进行词元分类微调，该实现通过 RoPE [33], 支持将输入长度扩展到 1024 个词元，从而解决 PDF 页面通常超过 512 个词元的问题。我们修改了 ERNIE Layout 的分类头，将词元分类为保留或删除，把目标从 DLA 模型通常采用的较广泛版面类别（例如 VILA 中最多有 10 类）简化为仅两类。

由于我们已针对词元分类对 ERNIE Layout 进行了微调，而每个文本片段可能包含多个词元，因此采用投票分类器 [7]来确定文本片段的预测结果。图 2 展示了一个可复制文本识别的示例。

3.2 候选项生成

在确定 PDF 中哪些文本片段可复制后，我们会合并相邻的可复制文本片段，同时将不相邻的可复制文本片段分开。所有合并后的文本片段都会存入列表，以供后续处理。

在候选生成过程中，CLD 会从列表中的第一个文本片段开始逐一检查，以找到匹配的 n 元语法。如果第一个文本片段中没有找到匹配项，

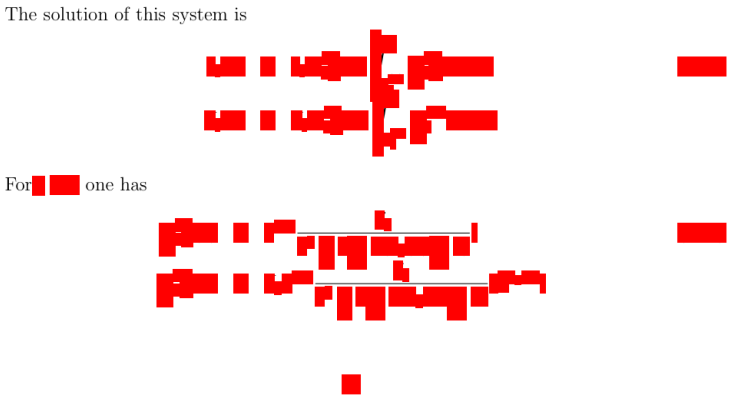


图2：可复制文本识别示例。在此示例中，我们的模型检测到与图1中相同的科学文档。删除部分（包括行内和独立的数学公式以及页码）会被识别并屏蔽。包含可复制文本的保留部分则由 CLD 的候选生成器返回。

则会继续搜索后续文本片段，直到找到匹配项或检查完所有文本片段为止。当在任一文本片段中检测到匹配的 n 元语法时，会将该片段中后续的词元作为候选项。

为了与 PDF 到 Markdown 转换遵循阅读顺序的特性保持一致，我们实现了一种截断机制。提取候选项后，PDF 到 Markdown 转换模型会评估这些候选项是否为序列中可接受的词元。如果某个候选项被接受，则表明模型正在处理包含该候选项的文本片段。此时，我们会重新组装文本片段列表，将当前活动文本片段及其后的所有文本片段移至列表前端，同时将此前的所有文本片段移至列表末尾。这样，CLD 就能够动态生成与 PDF 到 Markdown 转换当前进度同步的候选项，从而提高候选项的准确性和效率。

4 实验

4.1 PDF 到 Markdown 模型

我们使用两个公开可用的 PDF 到 Markdown 专用转换模型，以及两个通用、开源的视觉语言模型，在第 4.2 节所述的三个测试数据集上测试从头生成、mPLD 和 CLD 的转换速度与质量差异。所有测试均在配备 1×A100 GPU的机器上进行。

– 任务专用模型： Nougat[4] 和 Kosmos-2。5[5]**– 通用视觉语言模型：** Llama-3.2-11B-Vision-Instruct[6] 和 Qwen2- VL-72B-Instruct-AWQ[7]。由于量化技术 [21]，模型卡中将 Qwen2-VL 标注为 12.6B 参数。我们对这些模型使用提示词 1 中的 Prompt。

```
1 将以下 PDF 页面转换为 Markdown。仅返回Markdown，不要附加解释性文
```

提示 1：用于指示 Llama 和 Qwen2-VL 完成 PDF 到 Markdown 转换任务的提示。

4.2 数据集构建

由于目前尚无以 PDF 文件形式发布的现成数据集，我们从 arXiv 下载了 2023 年7月和8月 论文的 LaTeX 源代码包。随后，我们使用 LaTeX-Rainbow [8], , 它能够将 LaTeX 编译为 PDF，并为页面上的每个元素标注语义标签、阅读顺序，以及与数学公式和表格对应的 LaTeX 代码。总计标注了 14,320 篇论文。

从这些页面中提取文本片段，并根据上一步的标注结果将其标记为保留或删除。对于难以阅读的页面，例如包含整页图像、长表格或参考文献的页面，则从数据集中排除。最终构建的数据集包含 180,146 个页面，每个页面均标注了片段级文本可复制标签及其对应的边界框。我们随机将数据集划分为包含 162,127 个页面的训练集和包含 18,019 个页面的测试集。该数据集中的绝大多数页面均为英文。

我们的方法从 PDF 文件中提取文本片段，这需要访问学术论文的全文。由于 arXiv 不允许转载全文[8], ，我们发布了用于创建数据集的脚本以及相关文档的 arXiv 编号，以确保结果可复现。

[4] https://huggingface.co/facebook/nougat-base[5]
https://huggingface.co/microsoft/kosmos-2.5[6]
https://huggingface.co/meta-llama/Llama-3.2-11B-Vision[7]
https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct-AWQ[8]
https://info.arxiv.org/help/license/reuse.html\#full_文本

在 PDF 到 Markdown 转换速度测试中，我们从测试集中随机选取了 128 页。此外，我们还分别从经济学和量子物理学类别的论文中随机选取了 128 页，以测试 mPLD 和 CLD 在不同领域文档中的性能。

5 结果

5.1 可复制文本识别

我们使用 F1 分数作为度量指标，不进行平均，因为我们的模型是一个二元分类器。我们分别测试了词元级分数和片段级分数，并通过投票实现文本片段的预测，如第 3.1 节所示。

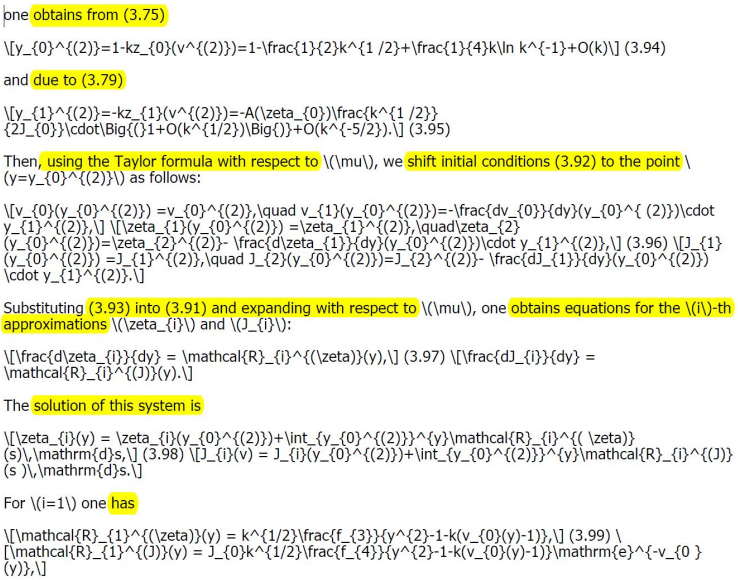
词元级的 F1 分数为**0.985**，而片段级的 F1 分数为**0.988**。分类器每页的平均推理时间为**0.03** 秒。

5.2 PDF 到 Markdown 的转换速度

表 1：我们在三个测试数据集上评估了 PDF 转 Markdown 转换模型和辅助生成方法的速度，其中 $t(s)$ 表示每页的平均生成时间，单位为秒。CLD 额外耗费的 0.03 秒源于可复制文本识别的开销。

		测试数据集					
		arXiv		经济学		量子物理	
骨干模型	辅助生成	$t(s)$	加速	$t(s)$	加速	$t(s)$	加速
Nougat-base（349M）	-	6.48	-	5.73	-	6.80	-
	mPLD	5.92	1.09×	5.05	1.13×	6.13	1.11×
	CLD	5.85 (+0.03)	**1.10**×	4.60 (+0.03)	**1.24**×	5.56（+0.03）	**1.22**×
Kosmos-2.5（1.37B）	-	12.21	-	11.09	-	12.77	-
	mPLD	9.42	1.30×	8.04	1.38×	9.23	1.38×
	CLD	8.37 (+0.03)	**1.45**×	7.60 (+0.03)	**1.45**×	7.85 (+0.03)	**1.62**×
Llama-3.2-Vision (10.6B)	-	33.83	-	30.53	-	35.03	-
	mPLD	29.20	1.16×	23.28	1.31×	32.08	1.09×
	CLD	26.02（+0.03）	**1.30**×	21.83（+0.03）	**1.40**×	27.61（+0.03）	**1.27**×
Qwen2-VL（12.6B）	-	138.85	-	119.42	-	156.06	-
	mPLD	97.65	1.32×	75.21	1.59×	100.00	1.56×
	CLD	91.53（+0.03）	**1.41**×	72.88（+0.03）	**1.64**×	91.88（+0.03）	**1.70**×

PDF 到 Markdown 转换速度测试结果如表 1 所示。在我们的实验中，mPLD 和 CLD 均加速了 PDF 到 Markdown 转换。就转换速度而言，模型越大，从 mPLD 和 CLD 中获益越多。CLD 中可复制文本识别模块的推理时间计入每页的平均转换时间。图 3 展示了 CLD 如何加速 Kosmos-2.5 的一个示例。



one obtains from (3.75)

\[y_{0}^{(2)}=1-kz_{0}(v^{(2)})=1-\frac{1}{2}k^{1 /2}+\frac{1}{4}k\ln k^{-1}+O(k)\] (3.94)

and due to (3.79)

\[y_{1}^{(2)}=-kz_{1}(v^{(2)})=-A(\zeta_{0})\frac{k^{1 /2}}
{2J_{0}}\cdot\Big{(}1+O(k^{1/2})\Big{)}+O(k^{-5/2}).\] (3.95)

Then, using the Taylor formula with respect to \(\mu\), we shift initial conditions (3.92) to the point \
(y=y_{0}^{(2)}\) as follows:

\[v_{0}(y_{0}^{(2)}) =v_{0}^{(2)},\quad v_{1}(y_{0}^{(2)})=-\frac{dv_{0}}{dy}(y_{0}^{ (2)})\cdot
y_{1}^{(2)},\] \[\zeta_{1}(y_{0}^{(2)}) =\zeta_{1}^{(2)},\quad\zeta_{2}
(y_{0}^{(2)})=\zeta_{2}^{(2)}- \frac{d\zeta_{1}}{dy}(y_{0}^{(2)})\cdot y_{1}^{(2)},\] (3.96) \[J_{1}
(y_{0}^{(2)}) =J_{1}^{(2)},\quad J_{2}(y_{0}^{(2)})=J_{2}^{(2)}- \frac{dJ_{1}}{dy}(y_{0}^{(2)})
\cdot y_{1}^{(2)}.\]

Substituting (3.93) into (3.91) and expanding with respect to \(\mu\), one obtains equations for the \(i\)-th approximations \(\zeta_{i}\) and \(J_{i}\):

\[\frac{d\zeta_{i}}{dy} = \mathcal{R}_{i}^{(\zeta)}(y),\] (3.97) \[\frac{dJ_{i}}{dy} =
\mathcal{R}_{i}^{(J)}(y).\]

The solution of this system is

\[\zeta_{i}(y) = \zeta_{i}(y_{0}^{(2)})+\int_{y_{0}^{(2)}}^{y}\mathcal{R}_{i}^{(\zeta)}
(s)\,\mathrm{d}s,\] (3.98) \[J_{i}(v) = J_{i}(y_{0}^{(2)})+\int_{y_{0}^{(2)}}^{y}\mathcal{R}_{i}^{(J)}
(s)\,\mathrm{d}s.\]

For \(i=1\) one has

\[\mathcal{R}_{1}^{(\zeta)}(y) = k^{1/2}\frac{f_{3}}{y^{2}-1-k(v_{0}(y)-1)},\] (3.99) \
[\mathcal{R}_{1}^{(J)}(y) = J_{0}k^{1/2}\frac{f_{4}}{y^{2}-1-k(v_{0}(y)-1)}\mathrm{e}^{-v_{0 }
(y)},\]

图 3：Kosmos-2.5 在 CLD 的辅助下进行 PDF 到 Markdown 转换的示例。在此示例中，高亮部分是由 CLD 提议并经 Kosmos-2.5 的验证步骤接受的词元。

表 2：消融研究结果，展示可复制文本识别（CTI）和候选生成（CG）对推理时间（t秒）的影响。两个模块均有助于缩短推理时间，二者结合可实现最高加速比。移除 CTI 和 CG 后，CLD 会退化为原始的 mPLD 方法。

	骨干模型							
	Nougat-base		Kosmos-2.5		Llama-3.2-Vision		Qwen2-VL	
模块	$t(s)$	加速	$t(s)$	加速	$t(s)$	加速	$t(s)$	加速
基线（mPLD）	6.13	-	9.23	-	32.08	-	100.00	-
基线 + CTI	5.80（+0.03）	1.05×	8.66（+0.03）	1.06×	27.83 (+0.03)	1.15×	92.83 (+0.03)	1.07×
base + CG	5.78	1.06×	8.68	1.06×	29.30	1.09×	96.99	1.03×
基准 + CTI + CG	**5.60 (+0.03)**	**1.10×**	**7.85 (+0.03)**	**1.17×**	**27.61 (+0.03)**	**1.16×**	**91.88 (+0.03)**	**1.08×**

表 3：辅助生成方法可无损保留原始输出。

辅助生成质量损失		
模型	mPLD	CLD
Nougat-base	0%	0%
Kosmos-2.5	0%	0%
Llama3.2-Vision	0%	0%
Qwen2-VL	0%	0%

5.3 消融研究

表 2 展示了不同模块组合对 Nougat-base 和 Kosmos-2.5 骨干模型推理时间及相对加速比的影响。此外，对 Llama 和 Qwen2-VL 进行的消融研究也显示出相同趋势。这些发现凸显了我们方法的实用价值。

5.4 PDF 到 Markdown 转换质量

除了转换速度外，我们还讨论了模型的转换质量。在这项工作中，对各个模型的准确率进行比较极为不可行，主要有以下原因：

– 辅助生成是一种无损方法，不会改变原始输出。因此，我们可以引用模型原论文中报告的准确率。– Nougat 和 Kosmos-2.5 均未公开发布其测试数据集，这可能是出于第 4.2 节所述的版权原因。因此，我们难以公平地复现其结果。– PDF 到 Markdown 转换任务的输出样式具有灵活性。具体而言，页面中浮动元素（包括表格和图形）的阅读顺序并不固定；Nougat 倾向于将浮动元素放在页面末尾，而 Kosmos-2.5 则倾向于将其放在页面开头。– 表格和公式的格式也具有灵活性。具体而言，Nougat 使用 LaTeX 格式表示表格，而 Kosmos-2.5 使用 Markdown 格式。公式的起始和结束符可以是美元符号，也可以是带括号的反斜杠。

因此，我们只能对转换质量进行极不公平的比较。毫无疑问，如表 3 所示，辅助生成不会降低模型生成的质量。我们采用了 Nougat [4] 中的评估指标。我们在测试数据集上对四个模型进行了评估，结果如表 4 所示。不过，分数较低并不一定意味着人工检查时可读性较差，因此这项评估并不能完美地衡量质量。

表 4：不同模型的 PDF 转 Markdown 转换质量。这并非公平比较，我们只是展示结果。

模型	编辑距离↓	BLEU↑	F1↑
Nougat-base	0.38	0.54	0.73
Kosmos-2.5	0.43	0.51	0.68
Llama-3.2-Vision	0.53	0.42	0.61
Qwen2-VL	0.46	**0.47**	**0.64**
Marker（基于流水线）	0.45	0.44	0.62

逐项分析发现，Llama 和 Qwen2-VL 难以准确遵循指令。Llama 经常在公式中生成错误的 LaTeX 代码，导致这些公式在 Markdown 中无法正确呈现。此外，它往往会将 PDF 内容重新构建为 LaTeX，而不是转换为 Markdown。Qwen2-VL 和 Llama 都会错放脚注，将它们置于文档末尾，而非引用位置，从而增加错误率。我们计划在未来的工作中改进评估指标。

6 结论

本文介绍了 mPLD 和 CLD，这两种轻量级辅助生成方法利用解码器 Transformer 模型实现端到端 PDF 转 Markdown 转换。mPLD 无需添加任何需要训练的参数，并且可以无缝集成到现有的 Transformer 解码器模型中。CLD 基于现成的 ERNIE Layout 模型，仅进行了少量微调，权重文件大小也保持在 27 MB。

PDF 出版物的数量逐年不断增加，且看不到放缓的迹象。基于流水线的方法有助于应对海量科学信息和新知识，但在实践中，这类技术的开发非常复杂，阻碍了基础设施和系统的构建，而这些基础设施和系统可用于追踪研究，并通过专门的科学搜索引擎和推荐系统等应用为科学界提供支持。

我们希望通过这些努力，使端到端 PDF 转 Markdown 转换的结构更加简单，从而更好地促进科学信息的无障碍访问和结构化提取。

7 局限性

在这项工作中，我们观察到，CLD 相较于 mPLD 的改进主要体现在：通过过滤机制和一种能够更好识别当前 PDF 到 Markdown 转换进度的停止机制，更准确地提取候选项。

然而，CLD 目前尚未显式检测解码器 Transformer 在页面上的转换进度位置。同期的工作 LOCR 通过使用附加在 Nougat 上的微型提示模块，预测 PDF 页面上下一个标记的位置，这或许能够增强候选项生成。下一步，我们计划将 LOCR 与 CLD 结合起来。

Kosmos-2.5、Nougat、Llama 和 Qwen2-VL 会舍弃页面中的图形，因为它们只能生成文本。ERNIELayout 能够识别这些图形。我们的目标是探索其他方法，将图形妥善整合到 Markdown 输出中。例如，Mistral OCR 会使用 Base64 字符串对图像进行编码。

致谢

本研究在 InsightsNet（insightsnet. org）研究项目框架内开展，该项目由德国联邦教育与研究部（BMBF）通过资助编号 01UG2130A 提供经费。我们衷心感谢 Wolf- gang Stille 博士以及 hessian.AI服务中心（由德国联邦教育与研究部 BMBF 通过资助编号 01IS22091 提供经费）和 hessian.AI创新实验室（由黑森州数字战略与创新部通过资助编号 S-DIW04/0013/003 提供经费）给予的支持。我们感谢 Sabine Bartsch博士对本研究的指导。我们还感谢匿名审稿人提出的深刻意见和建议，这些意见和建议极大地完善了本文。

参考文献

1. GROBID。https://github.com/kermitt2/grobid（2008–2024） 2. getomni ai：zerox。https://github.com/getomni-ai/zerox（2024），GitHub 仓库 3 21783 4 3 2024 2407。AI, M.：Llama 模型家族（），https://arxiv.org/abs/。。。Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.：Nougat：面向学术文档的神经光学理解（2024），https://openreview.net/forum?id=fUtxNAKpdV 5. Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J.D., Chen, D., Dao, T.：Medusa：采用多个解码头的简单大语言模型推理加速框架。载于：Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F.（编），《第41届国际机器学习会议论文集》。《机器学习研究论文集》，第235卷，第5209–5235页。PMLR（2024年7月21–27日），https://proceedings.mlr.press/v235/cai24b.html 6. Contributors, M.：MinerU：一站式开源高质量数据提取工具。https://github.com/opendatalab/MinerU（2024） 7. Diem, M., Kleber, F., Sablatnig, R.：论文片段的文本分类与文档版面分析。载于：2011年国际文档分析与识别会议。第854 175 8 858页，2011 10 1109 2011 –（）。https://doi.org/ . /ICDAR. . . Duan, C., Tan, Z., Bartsch, S.：LaTeX rainbow：通用 LaTeX 到 PDF 文档语义与版面标注框架。载于：Ghosal, T., Grezes, F., Allen, T., Lockhart, K., Accomazzi, A., Blanco-Cuaresma, S.（编），《第二届科学出版物信息提取研讨会论文集》。第56–67页。计算语言学协会，印度尼西亚巴厘岛（2023年11月）。https://doi.org/10.18653/v1/2023.wiesp-1.8，https://aclanthology.org/2023.wiesp-1.8/ 9. Fu, D., Guo, R., Khalighinejad, G., Liu, O., Dhingra, B., Yogatama, D., Jia, R., Neiswanger, W.：Isobench：对同构表示上的多模态基础模型进行基准测试。载于：首届语言建模会议（2024），https://openreview.net/forum?id

= 1 1 KZd EErRJ。Fu, Y., Bailis, P., Stoica, I., Zhang, H.：利用前瞻解码打破大语言模型推理的序列依赖。载于：Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F.（编），《第41届国际机器学习会议论文集》。《机器学习研究论文集》，第235卷，第14060–14079页。PMLR（2024年7月21–27日），https://proceedings.mlr.press/v235/fu24a.html

11. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.：LoRA：大语言模型的低秩适配。载于：国际学习表征会议（2022），https://openreview.net/forum?id=nZeVKeeFYf9 12. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.：LayoutLMv3：通过统一的文本与图像掩码进行文档人工智能预训练。载于：第30届ACM国际多媒体会议论文集。第4083–4091页。MM '22，计算机协会，纽约，纽约州，美国（2022）。https://doi.org/10.1145/3503161.3548112, https://doi.org/10.1145/3503161.3548112 13. Jaume, G., Kemal Ekenel, H., Thiran, J.P.：FUNSD：含噪扫描文档表单理解数据集。载于：2019年国际文档分析与识别研讨会（ICDARW）。第2卷，第1–6页（2019）。https://doi.org/10.1109/ICDARW.2019.10029 14. Joao Gante：辅助生成：迈向低延迟文本生成的新方向（2023）。https://doi.org/10.57967/hf/0638, https://huggingface.co/blog/assisted-generation 15. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.：无OCR文档理解变换器。载于：欧洲计算机视觉会议（ECCV）（2022） 16. Leviathan, Y., Kalman, M., Matias, Y.：通过推测解码实现变换器快速推理。载于：第40届国际机器学习会议论文集。ICML'23，JMLR.org（2023） 17. Li, K., Wigington, C., Tensmeyer, C., Zhao, H., Barmpalios, N., Morariu, V.I., Manjunatha, V., Sun, T., Fu, Y.：跨领域文档对象检测：基准测试套件与方法。载于：2020年IEEE/CVF计算机视觉与模式识别会议（CVPR）。第12912–12921页（2020）。https://doi.org/10.1109/CVPR42600.2020.01293 18. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.：DocBank：文档版面分析基准数据集。载于：Scott, D., Bel, N., Zong, C.（编），《第28届国际计算语言学会议论文集》。第949–960页。国际计算语言学委员会，西班牙巴塞罗那（线上）（2020年12月）。https://doi.org/10.18653/v1/2020.coling-main.82, https://aclanthology.org/2020.coling-main.82/ 19. Li, Y., Wei, F., Zhang, C., Zhang, H.：EAGLE-2：利用动态草稿树加速语言模型推理。载于：Al-Onaizan, Y., Bansal, M., Chen, Y.N.（编），《2024年自然语言处理实证方法会议论文集》。第7421–7432页。计算语言学协会，美国佛罗里达州迈阿密（2024年11月）。https://doi.org/10.18653/v1/2024.emnlp-main.422, https://aclanthology.org/2024.emnlp-main.422 20. Li, Y., Wei, F., Zhang, C., Zhang, H.：EAGLE：推测采样需要重新审视特征不确定性。载于：Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F.（编），《第41届国际机器学习会议论文集》。《机器学习研究论文集》，第235卷，第28935–28948页。PMLR（2024年7月21–27日），https://proceedings.mlr.press/v235/li24bt.html 21. Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.M., Wang, W.C., Xiao, G., Dang, X., Gan, C., Han, S.：AWQ：面向大语言模型压缩与加速的激活感知权重量化。载于：MLSys（2024）

22. Lv, T., Huang, Y., Chen, J., Cui, L., Ma, S., Chang, Y., Huang, S., Wang, W., Dong, L., Luo, W., 等：Kosmos-2.5：一种多模态文档理解模型。arXiv 预印本 arXiv:2309.11419（2023）23. Mathpix：mathpix-markdown-it。https://github.com/Mathpix/ mathpix-markdown-it（2024），GitHub 仓库 24. Mghari, H.E.：llama-ocr。https://github.com/Nutlope/llama-ocr（2024），GitHub 仓库 25. Mindee：doctr：文档文本识别。https://github.com/mindee/doctr（2021）26. NormXU：Ernie-layout-pytorch。https://github.com/NormXU/ERNIE-Layout-Pytorch（2023），GitHub 仓库 27. Peng, Q., Pan, Y., Wang, W., Luo, B., Zhang, Z., Huang, Z., Cao, Y., Yin, W., Chen, Y., Zhang, Y., Feng, S., Sun, Y., Tian, H., Wu, H., Wang, H.：ERNIE-layout：通过版面知识增强预训练，实现视觉丰富文档理解。载于：Goldberg, Y., Kozareva, Z., Zhang, Y.（编）计算语言学协会论文集：EMNLP 2022。第 3744–3756 页。计算语言学协会，阿布扎比，阿联酋（2022 年 12 月）。https://doi.org/10.18653/v1/2022.findings-emnlp.274，https: // aclanthology.org/2022.findings-emnlp.274/ 28. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.：Doclaynet：用于文档版面分割的大型人工标注数据集。载于：第 28 届 ACM SIGKDD 知识发现与数据挖掘会议论文集。第 3743–3751 页。KDD '22，计算机协会，美国纽约州纽约市（2022）。https://doi.org/10.1145/3534678.3539043，https://doi.org/ 10.1145/3534678.3539043 29. pymupdf：Pymupdf。https://github.com/pymupdf/PyMuPDF（2024），GitHub 仓库 30. Quattrini, F., Zaccagnino, C., Cascianelli, S., Righi, L., Cucchiara, R.：mugat：通过提供多页上下文改进单页文档解析（2024）31. Saxena, A.：提示查找解码（2023 年 11 月），https://github.com/ apoorvumang/prompt-lookup-decoding/ 32. Shen, Z., Lo, K., Wang, L.L., Kuehl, B., Weld, D.S., Downey, D.：VILA：利用视觉版面组改进科学PDF文档的结构化内容提取。计算语言学协会会刊**10**，376–392（2022）。https://doi.org/10.1162/tacl_a_00466，https://aclanthology.org/ 2022.tacl-1.22/ 33. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.：Roformer：采用旋转位置嵌入的增强型 Transformer（2023），https://arxiv.org/abs/2104. 09864 34. Summed-AI：parallex。https://github.com/Summed-AI/parallex（2024），GitHub 仓库 35. Sun, Y., Zhou, D., Lin, C., He, C., Ouyang, W., Zhong, H.S.：LOCR：用于光学字符识别的位置引导型 Transformer。载于：Al-Onaizan, Y., Bansal, M., Chen, Y.N.（编）计算语言学协会论文集：EMNLP 2024。第 5480–5497 页。计算语言学协会，美国佛罗里达州迈阿密（2024 年 11 月）。https://doi.org/10.18653/v1/2024.findings-emnlp.314，https://aclanthology.org/2024 314 36 .findings-emnlp. . Team, D.S.：Docling 技术报告。技术报告（2024 年 8 月）。https://doi.org/10 09869 37 48550 2408 09869，2408 2025 . /arXiv. . https://arxiv.org/abs/ . . VikParuchuri：marker。https://github.com/VikParuchuri/marker（），GitHub 仓库

38. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl：在任意分辨率下增强视觉语言模型对世界的感知（2024 12191 39 2409），https://arxiv.org/abs/ . . Wu, X., Hu, Z., Du, X., Yang, J., He, L.：通过动态残差特征融合进行文档版面分析。载于：2021 IEEE多媒体与博览会国际会议（ICME）。第 1 9428465 40 6 2021 10 1109 51207 2021 – 页（）。https://doi.org/ . /ICME . . . Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.：LayoutLMv2：用于视觉丰富文档理解的多模态预训练。载于：Zong, C., Xia, F., Li, W., Navigli, R.（编）第 59 届计算语言学协会年会暨第 11 届自然语言处理国际联合会议论文集（第 1 卷：长篇论文）。第 2579–2591 页。计算语言学协会，线上（2021 年 8 月）。https://doi.org/10.18653/v1/2021.acl-long.201, https://aclanthology.org/2021.acl-long.201/ 41. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.：Layoutlm：面向文档图像理解的文本与版面预训练。载于：第 26 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集。第 1192–1200 页。KDD '20，计算机协会，美国纽约州纽约市（2020）。https://doi.org/10.1145/3394486.3403172, https://doi.org/ 10.1145/3394486.3403172 42. Yang, H., Hsu, W.H.：使用循环卷积神经网络从科学文献中进行基于视觉的版面检测。载于：2020 年第 25 届国际模式识别会议（ICPR）。第 6455–6462 页（2021）。https://doi.org/10 9412557 43 1109 48806 2021 . /ICPR . . . Yang, N., Ge, T., Wang, L., Jiao, B., Jiang, D., Yang, L., Majumder, R., Wei, F.：基于参考的推理：大语言模型的无损加速（2023），https://arxiv.org/abs/2304 04487 44 . . Zhao, Y., Xie, Z., Liang, C., Zhuang, C., Gu, J.: Lookahead：一种用于大语言模型的推理加速框架，可实现无损生成准确率。载于：第 30 届 ACM SIGKDD 知识发现与数据挖掘会议论文集。第 6344–6355 页。KDD '24，计算机协会，美国纽约州纽约市（2024）。https://doi.org/10.1145/3637528.3671614, https://doi.org/10 3671614 45 1145 3637528 . / . . Zhong, X., Tang, J., Jimeno Yepes, A.：Publaynet：迄今最大的文档版面分析数据集。载于：2019 年国际文档分析与识别会议（ICDAR）。第 1015–1022 页（2019）。https://doi.org/10 00166 46 1109 2019 . / ICDAR. . . Zhu, W., Sokhandan, N., Yang, G., Martin, S., Sathyanarayana, S.: Docbed：面向复杂版面的多阶段OCR解决方案。AAAI人工智能会议论文集**36**(11), 12643–12649（2022 年 6 月）。https://doi.org/10.1609/aaai.v36i11.21539, https://ojs.aaai.org/index. php/AAAI/article/view/21539