当墨水遇见像素：面向人工校正金融文档的意图感知 OCR 与视觉语言模型

Akshata A Bhat[1], Daman Deep Singh[1],Niyati Chhaya[1]Hyperbots Inc.

arXiv:2611.32446v1 [cs.CL] 24 Nov 2026

摘要

文档中的人工注释——例如箭头、插入符号、删除线和插入——表达了纠正意图，但现有的光学字符识别和文档理解系统却将这些标记视为视觉噪声，而非语义编辑操作。我们提出了 意图感知 *OCR*，其任务是通过联合推理原始文档内容与传达编辑意图的视觉意图线索（手写或数字标记），生成校正后的转录文本。为支持这一任务，我们推出了 DocIntent，这是一个包含 1,600 份金融文档的多模态数据集，其中标注了超过 5,700 个视觉意图线索；每个线索均配有考虑版面布局且解析了意图的 OCR 目标文本。我们进一步推出了 IntentBench，这是一个包含 216 份文档的基准测试集，旨在评估模型在多样化真实世界注释条件下的稳健性。借助 DocIntent，我们在校正后的转录文本上进行直接意图监督，微调了 DocIntentOCR——一套包含 30 亿和 70 亿参数视觉语言模型的模型系列。在 IntentBench 上，经过微调的 30 亿和 70 亿参数模型分别取得了 0.81 和 0.83 的 F1 分数，显著优于商业闭源模型（例如 GPT-4.1（0.626））和开源模型（例如 LLaMA-3.2-Vision-90B（0.50）），表明意图感知监督能够使紧凑的领域专用模型在该任务上显著超越规模大得多的通用系统。该工作确立了意图感知 OCR 这一新颖的人工参与式文档理解任务，并提供了数据集、基准测试集和模型等基础资源，以推动交互式文档处理领域的研究。

1 引言

当会计师在发票上的错误行项目上划删除线，或合规官使用脱字符号插入更正后的数字时，这些标记便构成了一个决定——一项有意修改文档的指令。在金融、法律和审计工作流程中，印刷文档和数字文档会经常经过人工处理，并积累编码了纠正意图的注释：删除、替换、插入和核验。这些视觉线索——删除线、箭头、插入符号和圈选数值——构成了一种丰富而非正式的语言，审阅者借此传达文档应如何

内容应当改变。这个问题的规模十分庞大：据估计，全球每年产生的5500亿张发票中（Koch, 2019年），只有不到三分之一能够在无需人工干预的情况下实现直通式处理（Bartolini, 2024年），因此大多数发票都必须经过人工审核流程，并在其中不断积累此类注释。

光学字符识别（OCR）系统，从 Tesseract 等传统引擎（Smith，2007年）到 LayoutLM（Xu et al.，2020年）、Donut（Kim etal.，2022年）和 DeepSeek-OCR（Wei et al.，2025年）等现代视觉语言模型（VLMs），都旨在提取文本，同时对视觉噪声保持不变。其结果造成了根本性的脱节：OCR 输出反映的是原始的、未标注文档，而不是人工意图中的版本，因为视觉线索被视为噪声。在实践中，这迫使组织依赖额外的人工校正，或依赖在视觉文档空间之外运行的脆弱的基于规则的后处理流程（Bar-rachina et al.，2009年）。

我们认为，这一局限性反映出一个缺失的抽象：现有的 OCR 形式化定义缺乏对人类意图进行建模的机制。标注（手写或数字标注）并非任意标记，而是视觉编辑命令：删除线表示删除，插入符表示插入，箭头表示移动。我们将其形式化为意图感知 *OCR*，这是一项文档分析任务，其转录文本同时以印刷内容和视觉意图线索为条件，并将这些线索解释为语义编辑操作。根据范围不同，这些操作可能涉及词元、字段、表格单元格、行或整列。

为实现对该任务的系统性研究，我们推出了 DocIntent，这是一个包含 1,600 份合成金融文档的多模态数据集，并额外加入了超过 5,700 个视觉意图线索；每个样本都配有考虑版面的、已消解意图的 OCR 目标。为进行贴近现实的评估，我们提出了 IntentBench，这是一个包含 216 份现实世界标注金融文档的基准测试集。借助这些资源，我们在直接意图监督下对 DocIntentOCR——一个由 3B 和 7B 模型组成的系列——进行微调，结果表明，这些紧凑的领域特定模型的 F1 分数最高可达 0.83，显著优于规模大得多的闭源和开源模型，包括 GPT-4.1（0.626）和 LLaMA-3.2-Vision-90B（0.50）。

总而言之，本研究的贡献如下：

- 我们将意图感知 *OCR*形式化为一项任务：通过解读视觉反馈线索（如箭头、删除线和插入）作为编辑操作，生成校正后的转录文本。
- 我们推出了 DocIntent（1,600 份文档，5,700+ 个意图线索）和 IntentBench（216 份真实世界文档），这是首个面向意图感知 OCR 的数据集和基准测试集。
- 我们通过直接监督对 DocIntentOCR 进行微调，结果表明，意图感知微调使紧凑型视觉语言模型在此任务上的表现超越了规模大得多的模型。

通过将手写批注视为可执行的编辑，而非噪声，本研究为文档理解开辟了一个新方向——在这一方向中，OCR 系统不再是人工审阅工作流中的旁观者，而是参与者。

2 相关工作

光学字符识别研究已从基于规则的引擎（Smith, 2007）发展到布局感知预训练模型（Xu et al., 2020; Huang et al., 2022）、端到端架构（Kim et al., 2022; Blecher et al., 2023），以及 GOT-OCR（Wei et al., 2024）和 DeepSeek-OCR（Wei et al., 2025）等统一模型。通用视觉语言模型（Bai et al., 2025）和文档专用架构（Liu et al., 2024）在 DocVQA（Mathew et al., 2021）等基准测试集上取得了优异结果。这些系统以文本保真度为优化目标——手写注释会被逐字转录，或被视为噪声而抑制，绝不会被解读为编辑操作。离线手写识别（Marti and Bunke, 2002; Gervaisetal., 2024年）和删除线移除系统（Heil et al., 2021年）同样将注释视为需要识别或移除的信号。

人在回路范式（例如主动学习（Settles, 2009）和交互式机器翻译（Barrachina 等人, 2009）通过迭代优化，在模型生成输出后纳入用户反馈。意图感知 OCR 则反其道而行之：纠正意图直接以视觉批注的形式嵌入输入中，模型必须从单张静态图像中还原这些意图，无需反复澄清。

3 意图感知 OCR：任务定义

本节将意图感知 OCR 形式化为一项文档分析任务，定义视觉意图线索的范围及其引发的编辑操作，并说明评估协议。

3.1 形式化定义

令 I 表示一张包含印刷文本、版面结构以及零个或多个手写或数字注释的文档图像。传统 OCR 系统将 I 映射为转录文本 T，忠实复现所有印刷内容。意图感知 OCR 则将 I 映射为意图解析后的转录文本O，以反映视觉批注所指定的修改。重要的是，该定义不依赖于 O 的序列化格式：无论输出是保留二维文档结构的版面感知转录文本、按线性阅读顺序呈现的文本，还是任何其他文本表示，该定义均适用。下文定义的编辑操作作用于逻辑文档区域（词元、单元格、行、列、表格），因此不受最终转录文本线性化方式的影响。

我们对其建模如下。令 T 表示印刷内容的基础转录文本，令$\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ 表示图像中存在的视觉意图线索集合。每个线索 h_i 均与以下内容相关：

1. **线索类型** c_i 取自表4 （附录 A），2. **结构范围** s_i，用于指定编辑的粒度（标记、字段、单元格、行、列或表格），以及 3. 可选的**替换值** v_i（用于替换和插入操作）。

每条提示 h_i 都会对基础转录文本执行一项确定性编辑操作$e_i = (c_i, s_i, v_i)$。

意图解析后的输出定义为按顺序应用所有编辑操作：

$$O = e_n \circ e_{n-1} \circ \cdots \circ e_1(T)$$

在我们的形式化定义中，每个编辑 e_i 都针对基础转录文本中一个空间上相互独立的区域（特定的词元跨度、单元格、行、列或表格）。当编辑范围两两不相交时 [1]，组合操作满足交换律，顺序无关紧要。原则上，范围可能发生重叠（例如，同一行中的行级删除与单元格级替换）；我们将此类情况视为分类体系之外的标注歧义，不纳入本分类体系的范围，并且这类情况不会出现在DocIntent或IntentBench中。

关键在于，模型无法直接观察 T 或 $\mathcal{H}$，而必须从原始图像 I中推断二者。因此，该任务要求模型同时完成以下工作：(1) 识别文档中的印刷内容；(2) 检测并分类视觉批注；(3) 确定每条批注的结构范围；(4) 应用由此推断出的编辑，生成最终输出。

处理分类体系之外的标注。 表 4（附录A）中的分类体系并未涵盖所有手写标注惯例（例如圈注、下划线标注或页边批注）；这些形式被有意排除，因为它们在结构上存在歧义——例如，圆圈可能根据上下文表示删除、强调或疑问——如果没有操作上明确无歧义的定义就将其纳入，会破坏监督的一致性。对于不匹配十种已定义线索类型中任何一种的标注，预期的（也是当前的）模型行为是 不执行操作：底层文本按原样保留，同时模型应在输出中标记未匹配区域，而不是默默猜测编辑。我们将在局限性章节中讨论由此引入的消歧挑战以及由此产生的覆盖范围限制。

3.2 编辑操作分类体系

我们定义了意图引发变换的三个高层类别：

- **删除：** 现有文本或结构化内容被标记为移除（例如，一条没有替换值的删除线）。
- **替换：** 现有内容已失效，并替换为更正后的金额（例如，删除线旁附有手写更正）。
- **插入：** 在指定位置引入新内容，而不删除现有文本（例如，带有手写文本的插入符）。

每个类别可在不同的结构粒度上运作——从单个词元到整行、整列或整张表格。一个关键特性是，同一视觉提示会依据上下文映射到不同操作：单独的删除线会引发删除，而附近带有手写金额的删除线则会引发替换。表 4（附录 A）列举了所有提示类型、范围和预期OCR行为，图2 （附录A）展示了每一种的真实标注示例。

3.3 示例说明



图1：DocIntent 中的一张带注释发票样例。图中呈现了三个意图线索。

图 1展示了来自DocIntent的一个样例任务。人工审阅员做了以下处理：(1) 划掉发票号码 #437346 并画了一支箭头，指向手写替换内容 246952 （词元级别的箭头指示的替换）；(2) 在项目描述中的单词“drop”下方标注插入符，并手写“gold”（词元级别的插入）；(3) 涂掉电话号码（词元／字段级别的删除）。标准光学字符识别会逐字转录文档，并忽略这些注释。意图感知 OCR 会将每条注释解读为一项编辑操作，并生成校正后的转录文本：用246952 替换原始发票号码，将描述扩展为“Gold m.o.p orbit drop*gold* earrings”，并省略电话号码。

[1]这对我们提出的数据集中的所有样本均成立。

3.4 输出格式

模型会生成一种版面感知的纯文本转录文本，按照二维空间网格对文档内容进行序列化，其中每个词都由其行和列索引定位，同时保留表格边界、行列关联和阅读顺序（Bhat 等人，2026年）。我们采用这种格式，是因为金融文档具有固有的结构：即使每个词元都被正确识别，金额出现在错误的行或列中，也会导致转录文本在语义上不正确。扁平化的阅读顺序输出虽然便于人类阅读，却会丢失下游键值提取和表格解析所依赖的空间关系。

3.5 评估协议

我们通过比较预测的意图解析转录文本与真实输出，使用词元级F1分数进行评估，该分数基于完整的意图解析输出转录文本与完整真实文本计算——遵循文档理解基准测试中既有的做法，例如 DocVQA（Mathew 等人，2021年）而不是仅针对提取的片段或提示区域进行评估。这很适合意图错误会产生连锁影响的任务：漏掉删除操作或替换值错误都会改变周围的转录文本，而不仅仅是标注的词元，因此所有输出词元都会被统一评分，无论它们对应的是意图提示还是未标注内容。由于单靠词元级 F1 无法区分序列级性能下降与级联错误（例如，提示分类错误导致后续词元错位），我们还报告词错误率（WER），即基于线性化的意图解析转录文本，按照标准 Levenshtein 方法计算的词错误率（越低越好）。F1 分数衡量词元级别的编辑操作正确性，而 WER 衡量序列级转录质量；两者结合，比单独使用任一指标更全面地反映意图感知 OCR 的性能。所有评估均在IntentBench上进行。

4 数据集

我们介绍两项互补资源：DocIntent，一个用于训练的大规模合成数据集，以及 IntentBench，一个用于评估的真实世界基准测试集。二者均围绕视觉意图线索分类体系构建（附录A）。表1总结了其主要特征。

表 1：DocIntent 和 IntentBench 概览。

统计数据	DocIntent	IntentBench
# 文档数	1,659	216
意图线索总计	5,641	758
每份文档平均线索数	3.40	3.03
每份文档平均词元数	157.51	175.79
覆盖的提示类型	10	9
电子版 (%)	74.8	88.4
扫描版 (%)	25.2	11.6
文档类型数	33	18

4.1 DocIntent：训练数据集

DocIntent 是一个包含 1,659 份金融文档的多模态数据集（发票、报表和结构化表单），并增补了 5,641 个涵盖所有分类体系类别的视觉意图线索。基础文档取自真实世界来源和合成来源的均衡组合（附录C），因此，真实文档提供了生产环境金融工作流程中的版面多样性，而合成文档则支持在特定结构粒度上确定性地放置意图线索。表1 中的 74.8% 数字渲染 / 25.2% 实体标注并扫描的划分，指的是意图线索的呈现方式（模拟叠加与实体手写），而不是底层基础文档的来源；我们在附录C中明确说明了这一点。表1总结了关键统计数据；按线索类型划分的明细见附录B。完整的标注协议详情请参阅附录C 。

监督信号。DocIntent 中的每份文档都配有经过意图解析且感知版面的转录文本，该文本应用所有视觉线索，同时保留二维文档结构（表格边界、行列关联、阅读顺序）。

4.2 IntentBench：评测基准

IntentBench 是一个独立构建、未参与 DocIntent 且完全排除在训练之外的留出基准。它包含多样的金融文档，在手写风格、标注密度和布局方面差异显著，用于测试模型在训练分布之外的泛化能力。真实标注为经意图解析且版面感知的转录文本。汇总统计见表1 ，详细分类见附录 B 。

表格 2：IntentBench 光学字符识别性能

模型	F1	WER
Gemini 2.5 Flash	0.6653	1.3306
Gemini 2.5 Pro	0.6351	0.6009
GPT-4.1	0.6261	0.9058
Claude Sonnet 4.6	0.5974	0.9308
Mistral-Large-3-675B	0.6077	0.9840
Llama-3.2-90B-Vision	0.4982	1.0986
Qwen3-VL-32B-Instruct	0.5414	1.2720
o3	0.7686	0.4869
GPT-4o	0.8052	0.4232
Qwen2.5-VL-7B-Instruct（基础版）	0.6485	0.8172
Qwen2.5-VL-3B-Instruct（基础版）	0.6563	0.7404
DocIntentOCR-3B	0.8102	0.5622
DocIntentOCR-7B	**0.8328**	**0.2589**

标注者间一致性。 两位标注者独立标注了全部 216 份 IntentBench 文档。一致性通过标准化 Levenshtein 比率（Levenshtein, 1966年）衡量，平均相似度为 0.77（$\sigma = 0.12$），表明一致性中等偏高。完整的标注者间一致性统计见附录 D。

5 实验

我们在 DocIntent 上对 Qwen2.5-VL-Instruct 模型的 3B 和 7B 变体进行微调，采用 80/20 的训练/验证划分。微调后的模型分别记为 DocIntentOCR-3B和 DocIntentOCR-7B，并在 IntentBench上进行评估。两个模型均使用 LoRA（r=16， α=32）对语言模型层进行微调，同时解冻最后一个视觉编码器块，并使用 AdamW（1×10^{-4}/1×10^{-5}）进行优化，分别训练 7 个和 5 个 epoch，使用两块 NVIDIA A100 40GB GPU。为公平评估零样本模型，我们提供了明确定义注释语义的提示（附录 E.1），以及超参数配置（附录 E.2）。

6 结果

表格2报告了 IntentBench F1分数。结果揭示了一个清晰的规律：在 DocIntent 上进行意图感知微调才是决定性因素，而非模型规模。零样本模型（包括 GPT-4.1（0.626））表现不佳，因为它们没有机制将视觉修订标记解读为语义编辑操作。 DocIntentOCR-3B（0.810）和 DocIntentOCR-7B（0.833）均优于所有基线，表明经过意图感知监督训练的紧凑型本地可部署模型，能够成为一种实用且注重隐私的替代方案。

表 3：在 IntentBench 上消融意图监督与模型规模对 F1 分数的影响。

消融实验	From	To	ΔF1
意图监督（3B）	0.656	0.810	+0.155
意图监督（7B）	0.649	0.833	+0.184
模型规模（受监督）	0.810	0.833	+0.023

，可替代大型前沿 API（参见图 3，其中与 GPT-4.1 进行了定性比较）。

消融实验。 表格3通过将微调前的 Qwen2.5-VL-Instruct 检查点作为无监督参照，并保持架构、提示词和解码设置一致，区分了意图感知监督与模型规模各自的贡献。意图监督在 3B 模型上带来 +00.155 的 F1 分数提升，在 7B 模型上带来 +00.184 的 F1 分数提升；相比之下，单纯增大模型规模（从 3B→7B，二者均接受意图监督）仅带来 +00.023 的 F1 分数提升（约 7×小得多）。这证实，在此任务中，意图感知监督而非参数数量才是性能提升的主要驱动因素。

分层误差分析。 仅凭词元级 F1 分数会混淆不同的失效模式：编辑区域中的意图解读错误与未编辑区域中的无端幻觉；此外，不同提示类型的表现可能存在显著差异。我们在附录 F中从这两个维度分解性能：（i）按提示类型划分的编辑精确匹配；（ii）编辑区域与非编辑区域的 F1 分数。 DocIntentOCR- 7B 在带替换内容的删除线上取得最高的编辑精确匹配；这是出现频率最高的提示类型（209 个实例，涉及 185 份文档）。两个微调变体在几乎所有提示类型上的表现均领先或具有竞争力，唯独在整张表格删除和箭头指示的替换上，o3 略占优势。 DocIntentOCR-3B 的编辑区域 F1 分数与 GPT-4o 相当（0.862 对 0.868），且在所有受评模型中取得最高的非编辑区域 F1 分数（0.832），表明它在未标注内容中产生的幻觉最少。

跨领域泛化。 DocIntent 和 IntentBench 均以金融文档为基础。为探究意图感知监督能否迁移到这一领域之外，我们在涵盖三个非金融领域（法律合同、医疗记录和房地产合同；每类 10 份）的 30 份文档上开展了初步的跨领域试点研究，并使用相同的模

型，根据 IntentBench 式的标注解读方式进行评估。完整结果见附录G。DocIntentOCR-3B 和 DocIntentOCR-7B 即使在领域外，其表现仍优于零样本视觉语言模型（F1 分数分别为 0.896 和 0.939，而最佳零样本基线o3的 F1 分数为 0.875），且各模型的相对排名在不同领域中基本保持不变。这表明，意图感知监督能够泛化到标注解读技能，而非过度拟合金融特定内容。我们强调，这只是一项初步试点研究（30 份文档，并非完整的基准测试集）；专门的跨领域基准测试集将留待未来工作，我们也会在局限性章节中进一步讨论这一局限。

7 结论与未来工作

我们提出了意图感知 OCR，这是一项新颖的文档理解任务，旨在将人工注释——删除线、箭头、插入符号和插入内容——解读为语义编辑操作。为支持这项任务，我们推出了 DocIntent，这是一个包含 1,600+份带注释的金融文档的数据集，以及 IntentBench，一个包含 216 份文档的真实世界基准测试集。在 DocIntent上进行微调后，DocIntentOCR 在 3B 和 7B 两种配置下均显著优于 GPT-4.1 及其他前沿模型（零样本设置），这表明意图感知监督是关键因素，而非模型规模。初步的跨领域试点研究（附录 G）进一步表明，这一优势可扩展到金融文档之外的法律、医疗保健和房地产合同领域。未来工作包括发布一个专门的跨领域基准测试集，以覆盖更广泛的文档范围；扩展分类体系，以处理组合式和歧义批注（例如圈注、下划线标注和页边批注）；以及通过合成图像生成来扩大训练数据规模。

局限性

我们的主要数据集和评估以金融文档为基础；尽管我们的跨领域试点研究（附录 G，涵盖法律、医疗保健和房地产领域的 30 份文档）初步展现了模型的泛化能力，但它无法替代完整的跨领域基准测试，扩大验证范围仍是未来的工作。该分类体系未涵盖所有注释风格，例如圈注、下划线和页边批注；由于这些注释在结构上存在歧义，因此未将其纳入（例如，圆圈根据上下文可表示删除、强调或疑问），且无法对其作出操作层面上毫无歧义的定义，因此模型目前会将未定义的注释类型视为无操作。将这些含糊惯例以及其他领域的注释风格（例如法律文件修订中表示接受/拒绝的语义）纳入分类体系，是一个重要的研究方向。以数字数据为主的训练数据（74.8%）可能会限制模型对多样化真实世界手写文本的鲁棒性。标注者间一致性为 0.77，表明空间位置模糊的批注存在主观解读空间；对一致性较低的样本对进行人工检查后发现，分歧主要源于空间位置模糊的批注放置方式（例如，插入符位于两个候选插入点之间，或修正内容靠近单元格边界），而非标注者对分类体系的理解存在系统性差异。尽管词元级 F1 分数和词错误率指标辅以按提示类型及已编辑/未编辑区域进行的分析（附录 F），但目前尚未对系统性与随机性失效模式进行完整的定性分类。最后，我们分类体系中的结构范围（标记、字段、单元格、行、列、表格）主要针对表格布局而设计；具有不规则或自由格式布局的文档可能需要采用其他范围定义。

伦理声明

文档DocIntent 和 IntentBench 均为合成生成，或来源于公开可用的文档集合和模板（附录 C）；未使用任何私人、非公开的财务记录或个人身份信息。人工标注员均为自愿招募，已获知任务目的，并因其工作时间得到公平报酬。标注说明经过设计，以避免主观性或有害内容。

本文部分内容是在大语言模型工具（Claude、OpenAI）的协助下起草的。所有科学内容、实验结果和结论均由作者完成并核验。

参考文献

Shuai Bai、Keqin Chen、Xuejing Liu 和另外 1 位作者。2025 年。Qwen2.5-VL 技术报告。*arXiv* 预印本 *arXiv:2502.13923*。

Sergio Barrachina、Oliver Bender 等1人。2009年。计算机辅助翻译的统计方法。计算语言学，35(1)：3–28。

Andrew Bartolini。2024年。电子应付账款的现状2024：资金永不眠。技术报告，Ardent Partners。

Akshata Bhat、Sindhu Naganna、Saiful Haq、Prayas Khatri、K.C.R. Tamataam、N. Arun、Niyati Chhaya 和 Pawan Bhattacharya。2026年。SAVIOR：面向光学字符识别表征的视觉语言模型样本高效适配。载于*IEEE/CVF* 冬季计算机视觉应用会议（*WACV*）研讨会论文集，第 719–728 页。

Lukas Blecher、Guilherme Correia 和 Klemens Müller。2023年。Nougat：面向学术文档的神经光学理解。*arXiv* 预印本 *arXiv:2308.13418*。Philippe Gervais、Asya Fadeeva 和 Andrii Maksai。2024年。MathWriting：用于手写数学表达式识别的数据集。*arXiv* 预印本 *arXiv:2404.10690*。Raphaela Heil、Ekta Vats 和 Anders Hast。2021年。使用循环生成对抗网络去除手写单词中的删除线。载于 *ICDAR* 论文集，第 572–586 页。Felix Hertlein、Alexander Naumann 和 Patrick Philipp。2023年。Inv3d：用于模板引导单图像文档去扭曲的高分辨率三维发票数据集。《国际文档分析与识别杂志》（*IJDAR*），第 719–728 页。Zhe Huang、Jianlong Wu、Yibo Chen 及其他 1 人。2022年。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* 多媒体国际会议论文集，第 4083–4091 页。Geewook Kim、Teakgyun Kim、Jinyoung Kim 和 Byung-Seok Roh。2022年。Donut：无需光学字符识别的文档理解变换器。载于欧洲计算机视觉会议（*ECCV*）论文集，第 414–430 页。Bruno Koch。2019年。2019–2025年电子发票之旅。技术报告，Billentis。Vladimir I. Levenshtein。1966年。能够纠正删除、插入和逆序的二进制编码。《苏联物理学文献》，10(8)：707–710。Yuliang Liu、Biao Yang、Qiang Liu 及其他 1 人。2024年。TextMonkey：用于理解文档的无 OCR 大型多模态模型。*arXiv* 预印本 *arXiv:2403.04473*。Urs-Viktor Marti 和 Horst Bunke。2002年。IAM 数据库：用于离线手写识别的英语句子数据库。《国际文档分析与识别杂志》，5(1)：39–46。Minesh Mathew、Dimosthenis Karatzas 及其他 1 人。2021年。DocVQA：用于文档图像视觉问答的数据集。载于 *IEEE/CVF* 计算机视觉应用冬季会议（*WACV*）论文集，第 2200–2209 页。

Burr Settles。2009年。主动学习文献综述。技术报告，威斯康星大学麦迪逊分校。

Ray Smith。2007年。Tesseract OCR 引擎概述。载于国际文档分析与识别会议（*ICDAR*），第 629–633 页。

UCSF 行业文档库。2024年。UCSF 行业文档库。https://www.industrydocuments.ucsf.edu/。

Haoran Wei、Chenglong Liu、Jinyue Chen 和其他 1 位作者。2024年。通用 OCR 理论：通过统一的端到端模型迈向 OCR 2.0。*arXiv* 预印本 *arXiv:2409.01704*。Haoran Wei、Yaofeng Sun、Yukun Li 和其他 1 位作者。2025年。DeepSeek-OCR：上下文光学压缩。*arXiv* 预印本 *arXiv:2510.18234*。Yiheng Xu、Minghao Li、Lei Cui、Shaohan Huang、Furu Wei 和 Ming Zhou。2020年。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会议论文集，第 1192–1200 页。

意图线索分类体系

表格4 列出了完整的分类体系。虽然这十种提示类型源于金融文档，但三类编辑操作（删除、替换、插入）属于与领域无关的基本操作，而用于表达这些操作的视觉线索（删除线、插入符号、箭头）则是各专业领域通用的惯用标记。将覆盖范围扩展至法律文件修订标记、医疗记录或工程变更单，可能需要增加提示类型并定义范围；在非金融领域开展实证验证仍是未来研究的重要方向。

图2 展示了 DocIntent和 IntentBench 中实际出现的每种提示类型各一个代表性标注示例（表 1总结了这十种提示类型）。

表4：分类 面向意图感知 OCR 的视觉与语义意图线索分类法。每种线索都会在特定的结构粒度上对基础转录文本进行确定性的编辑。

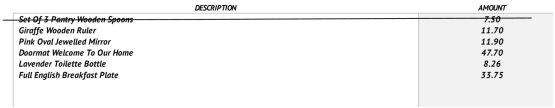
类别	线索类型	范围	预期光学字符识别行为
删除	删除线	标记 / 字段	标记内容不会出现在输出中。
	行删除线	表格行	删除线行中的所有单元格均已被移除。
	列删除线	表格列	删除线列中的所有金额均已已移除。
	整表删除线	表格	整个表格不会出现在输出中。
替换	删除线 +手写金额标记 / 字段		删除线文本被附近的手写金额替换。
	箭头指示的替换	标记 / 字段	删除线文本被箭头所指向的金额替换。
	单元格更正	表格单元格	单元格金额已替换为手写的更正内容·
	行更正	表格行	该行中的每个单元格都会替换为其各自的单元格替换值。
	列更正	表格列	该列中的每个单元格都会替换为其替换值。
插入	插入脱字符	标记 / 字段	新的手写文本插入在插入符位置，且不删除现有内容。
	箭头指示的添加	标记 / 字段	新内容插入在箭头所指示的位置由箭头标示。
	行内手写添加	标记 / 字段	新的手写内容插入在相邻位置，与现有文本相邻，且没有明确的插入符。
选择 / 验证	勾号（✓）	标记 / 字段	标记内容已被明确选择或确认为有效。
	叉号（X）	标记 / 字段	标记内容被明确拒绝或判定为无效，（通常意味着删除）。
	实心圆 / 高亮	标记 / 字段	圈出或高亮的内容会受到强调视相关情况或所选内容而定。
自然语言	指令覆盖	变量	自由格式文本（例如“替换为 500”）明确地覆盖现有内-显式覆盖现有内容。
	指令删除	变量	文本线索（例如“忽略这一行”）会删除指定的内容。
	指令添加	变量	文本线索（例如“添加 GST”）会引入新的内容，并放置在符合上下文的位置。



无替换删除线（删除）。

（空）

意图解析后的 OCR 输出。

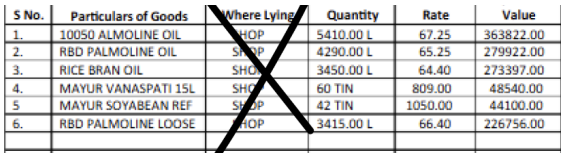


DESCRIPTION	AMOUNT
~~Set Of 3 Pantry Wooden Spoons~~	~~7.50~~
Giraffe Wooden Ruler	11.70
Pink Oval Jewelled Mirror	11.90
Doormat Welcome To Our Home	47.70
Lavender Toilette Bottle	8.26
Full English Breakfast Plate	33.75

表格行删除（删除）。

描述	数量
长颈鹿木尺	11.70
粉色椭圆宝石镜	11.90
欢迎回家门垫	47.70
薰衣草香水瓶	8.26
全英式早餐盘	33.75

意图解析后的 OCR 输出。

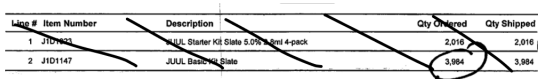


S No.	Particulars of Goods	Where Lying	Quantity	Rate	Value
1.	10050 ALMOLINE OIL	SHOP	5410.00 L	67.25	363822.00
2.	RBD PALMOLINE OIL	SHOP	4290.00 L	65.25	279922.00
3.	RICE BRAN OIL	SHOP	3450.00 L	64.40	273397.00
4.	MAYUR VANASPATI 15L	SHOP	60 TIN	809.00	48540.00
5	MAYUR SOYABEAN REF	SHOP	42 TIN	1050.00	44100.00
6.	RBD PALMOLINE LOOSE	SHOP	3415.00 L	66.40	226756.00

表格列删除（删除）。

序号 / 货物详情	数量	Rate	金额
1. 10050 阿尔莫莱恩油	5410.00 升	67.25	363822.00
2. 精炼漂白脱臭棕榈油	4290.00 升	65.25	279922.00
3. 米糠油	3450.00 升	64.40	273397.00
4. Mayur 人造黄油 15升	60 税务识别号	809.00	48540.00
5. Mayur 大豆精炼油	42 税务识别号	1050.00	44100.00
6. RBD 棕榈油散装	3415.00 升	66.40	226756.00

意图解析后的 OCR 输出。

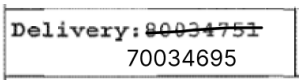


Line #	Item Number	Description	Qty Ordered	Qty Shipped
1	J1D1023	JUUL Starter Kit Slate 5.0% 4.2ml 4-pack	2,016	2,016
2	J1D1147	JUUL Basic Kit Slate	3,984	3,984

整表删除（删除）。

(empty)

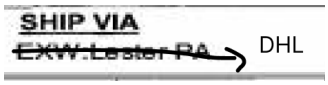
意图解析后的 OCR 输出。



删除线替换（替换）。

交付：70034695

意图解析后的 OCR 输出。



箭头指示的替换（替换）。

运输方式
DHL

意图解析后的 OCR 输出。

Date	Description	Withdrawals	Deposits	Balance
07/01/2023	Opening Balance	-	-	~~$5,000.00~~ **$5.200.00**

表格单元格替换 (替换)。

Date	描述	取款	存款	余额
2023年07月01日	期初余额	-	-	$5,200.00

意图解析后的 OCR 输出。

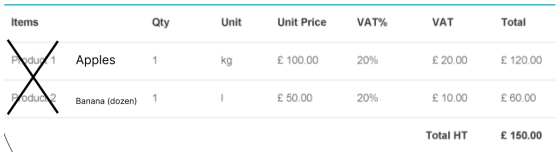


	Description	Ordered	Delivered	Outstanding
Item #	Ethernet Cable	10	10	0
Item #	SSD Storage	5	5	0
Item #	~~Monitor~~ Mouse pad	~~3~~ 4	~~3~~ 4	0

表格行替换（替换）。

	描述	已订购	已交付	未完成
商品编号	以太网线	10	10	0
商品编号	SSD存储	5	5	0
商品编号	鼠标垫	4	4	0

意图解析后的 OCR 输出。



Items		Qty	Unit	Unit Price	VAT%	VAT	Total
~~Product 1~~	Apples	1	kg	£ 100.00	20%	£ 20.00	£ 120.00
~~Product 2~~	Banana (dozen)	1	l	£ 50.00	20%	£ 10.00	£ 60.00
						Total HT	£ 150.00

表格列替换（替换）。

商品	Qty	Unit	单价	VAT%	总计
苹果	1	kg	£100.00	20%	£120.00
香蕉（打）	1	l	£50.00	20%	£60.00
			总计 HT：£150.00		

意图解析后的 OCR 输出。



插入脱字符（插入）。

第482号邮政信箱，第二街，威斯敏斯特大厦
米尔班克7号，伦敦 SW1P 3JE

意图解析后的 OCR 输出。

图2：DocIntent/IntentBench 中全部十种意图线索类型的代表性标注示例，9以及与之并列展示的字面意图解析后的 OCR 输出。

B 数据集统计

表格5 和6 分别报告了DocIntent 和IntentBench中视觉意图线索按线索类型的分布情况。在两个划分中，替换线索均占主导，其中带删除线的替换和箭头指示的替换是最常见的类型。

表格 5：DocIntent 中视觉意图线索的分布。每一行报告线索实例总数、包含至少一个实例的文档数量，以及在全部 1,659 个文档中的平均每文档频率。

类别	线索类型	总计	文档	平均/文档
删除	无替换删除线 986		833	0.59
	表格行删除	189	169	0.11
	表格列删除	333	303	0.20
	整表删除	133	130	0.08
替换	删除线/替换 1,880 1,233			1.13
	箭头指示的替换 1,214 1,010			0.73
	表格单元格替换	463	315	0.28
	表格行替换	16	12	0.01
	表格列替换	45	33	0.03
插入	插入脱字符	382	331	0.23
总计		**5,641**	–	**3.40**

表格 6：IntentBench 中视觉意图线索的分布。每行报告线索实例的总数、包含至少一个线索实例的文档数量，以及每个文档的平均频率。

类别	线索类型	总计	文档	平均/文档
删除	无替换删除线 160 143			0.74
	表格行删除	1	1	0.00
	表格列删除	30	30	0.14
	整表删除	12	12	0.06
替换	删除线替换	209	185	0.97
	箭头指示的替换	192	180	0.89
	表格单元格替换	28	23	0.13
	表格列替换	1	1	0.00
插入	插入脱字符	22	21	0.10
总计		**655**	–	**3.03**

C 标注协议

数据来源。 DocIntent 基础文档（注释前）来自涵盖 33 种文档类型的真实世界来源和合成来源的均衡组合。**真实世界文档** （约占语料库的 53%）来源于公开可用的 UCSF 行业文档库（UCSF 行业文档库，2024年）以及其他公开发布的金融文档集合，为生产环境金融工作流中所体现的版面和内容多样性作出了贡献。 **合成文档** （约占 47%）通过两条互补的流程生成：（i）使用定制构建的 HTML 金融模板，其中填充了合成生成的内容（姓名、行项目、数量），并经过设计以确保数学一致性（例如，正确的小计、税额和总计），随后通过无头浏览器渲染为图像；（ii）Inv3D 数据集（Hertlein 等人，2023年），该数据集提供高分辨率的合成三维渲染发票。合成文档能够精确控制版面、表格结构和空间组织（这些属性对于在特定结构粒度上确定性地放置意图线索至关重要），而真实世界部分则使数据集建立在真实的文档多样性之上。需要指出的是，表1 中报告的 74.8% 数字版 / 25.2% 扫描版划分，指的是叠加在这些基础文档上的意图线索的呈现方式（数字模拟的注释与实际手写并扫描的注释），而不是基础文档本身的来源。

手写批注由五名具有商科背景、且此前接触过会计和商业文档的标注者完成。研究人员向标注者提供了针对具体文档的更正说明，明确指出需要删除、替换或插入的内容，并要求他们在无法查看模型预测或自动建议的情况下，使用自然的手写方式进行更正。每份完成标注的文档均由第二名标注者独立复核，以确认所有预期提示均已正确应用且在视觉上清晰明确。标注者均获得了合理的报酬，且本研究遵循既定的伦理标准。

D 标注者间一致性

为评估IntentBench参考注释的可靠性，两位标注者分别为全部 216 份文档生成了意图解析后的转录文本；我们使用归一化莱文斯坦比率（Levenshtein，1966年）量化一致性。表7报告了由此得出的相似度统计数据，平均分为 0.77，表明标注者之间具有中等偏高的一致性。

表 7：使用模糊字符串相似度（归一化莱文斯坦比率）对 IntentBench 注释进行评估的标注者间一致性（IAA）。

指标	金额
注释对数量	216
平均相似度	0.77
最小相似度	0.31
最大相似度	0.99
标准差	0.12

E 实验配置

E.1 提示词

为评估DocIntentOCR-3B和DocIntentOCR-7B，我们采用与训练期间相同的指令：“你是一名 *OCR* 助手。分析文档图像并转录所有可见文本，将任何注释解释为编辑指令。生成最终的校正后的转录文本。”

对于所有其他未微调模型和闭源模型，我们使用稍作扩展的提示词，以明确界定注释语义：“你是一个用于处理带注释金融文档的 *OCR* 助手。请将注释解释为编辑指令（例如，用删除线表示删除或替换，用箭头表示替换，用插入符号表示插入，等等）。输出最终的校正后的转录文本，并保留版面。”

E.2 推理超参数

所有模型均采用温度 0.1、top-p 0.90 和 8,000 个词元的最大输出长度进行评估。零样本基线分别通过 OpenAI API（GPT-4.1）、Anthropic API（Claude Sonnet 4.6）和 AWS Bedrock（Llama-3.2-90B 与 Mistral-Large-3-675B）调用。

F 分层误差分析

F.1 按提示类型划分的编辑精确匹配

表 8 报告了编辑精确匹配（EEM）这一指标。与词元级 F1 分数相比，该指标更为严格：只有在正确的编辑标记出现且删除线标记的原始内容不存在时，才计为 1；表中还按提示类型对所有在IntentBench上评估的模型进行了细分。提示类型的归属依据 IntentBench 的真实标注标记确定。我们报告了出现频率最高的七种提示类型（其余罕见类型的样本量，例如表格行/列替换，太小，无法获得稳定的逐类型估计；参见表6）。

DocIntentOCR-7B 在带替换的删除线这一线索类型上取得最高的 EEM；该线索类型也是出现频率最高的类型（209 个实例，涉及 185 份文档）。DocIntentOCR-3B 则在不带替换的删除线上表现最佳。两种微调变体在表格列删除任务上均处于领先地位。在所有删除和替换线索类型中，微调模型均以显著优势稳定超越其各自的基础模型（Qwen2.5-VL-7B/3B）。除整表删除和箭头指示的替换外，微调模型在所有线索类型上均具有竞争力或处于领先地位；在前两类任务中，o3 略占优势（分别为 0.80 对 0.68–0.71，以及 0.77 对 0.73–0.75）。

F.2 编辑区域与非编辑区域分析

仅使用词元级 F1 分数会混淆编辑区域中的错误（意图解读错误或编辑应用错误）与非编辑区域中的错误（无端产生的幻觉或文本退化）。表格 9通过使用已知的提示边界框，将性能分解为编辑区域 F1 和非编辑区域 F1；这些提示边界框来自IntentBench 注释，用于将输出词元划分为编辑片段和非编辑片段，并分别计算 F1 分数。

DocIntentOCR-3B 在编辑区域 F1（0.8624 对 0.8677）方面与 GPT-4o 相当，并在所有评测模型中取得最高的非编辑区域 F1（0.8321），表明其在未注释内容中产生的幻觉最少。相比之下，若干强大的零样本基线模型（例如 GPT-4.1、Claude Sonnet 4.6、Qwen3-VL-32B）的编辑区域 F1 与非编辑区域 F1 之间存在明显差距，这表明它们的错误主要集中在未编辑内容中，而非源于错误应用编辑。

G 跨领域试点研究

DocIntent 和 IntentBench 均以金融文档为基础，主要结果并未直接提供跨领域迁移的证据。为初步评估泛化能力，我们针对 30 份文档开展了一项聚焦于跨领域的试点研究，涵盖三个非金融领域：法律合同（10 份）、医疗保健记录（10 份）和房地产合同（10 份），并按照与 IntentBench 相同的标注解读方式，对同一组模型进行评估（根据相同的分类体系和标注协议，在文档上叠加意图线索，具体协

表 8：IntentBench 上按线索类型划分的编辑精确匹配（EEM），针对七种最常见的线索类型。粗体标出每行的最佳分数。

线索类型	DocIntentOCR-7B	DocIntentOCR-3B	Qwen2.5-VL-7B	Qwen2.5-VL-3B	GPT-4.1	GPT-4o	o3	Claude-S.-4.6	Gemini-2.5-Flash	Qwen3-VL-32B	Gemini-2.5-Pro	Llama-3.2-90B	Mistral-L.-3
无替换删除线	0.83	**0.85**	0.67	0.79	0.73	0.83	0.81	0.71	0.81	0.67	0.83	0.77	0.73
表格列删除	**0.81**	**0.81**	0.69	0.67	0.69	**0.81**	0.71	0.70	0.72	0.59	0.69	0.44	0.69
整表删除	0.68	0.71	0.71	0.68	0.72	**0.80**	**0.80**	0.68	0.64	0.57	0.71	0.54	0.69
删除线替换	**0.79**	0.77	0.55	0.57	0.72	0.78	0.75	0.70	0.66	0.62	0.62	0.57	0.57
箭头指示的替换	0.73	0.75	0.53	0.44	0.66	0.76	**0.77**	0.69	0.68	0.61	0.64	0.37	0.50
表格单元格替换	0.88	0.94	0.88	0.59	0.88	**1.00**	**1.00**	**1.00**	0.88	0.94	0.71	0.88	0.82
插入脱字符	**1.00**	**1.00**	0.80	0.80	**1.00**	**1.00**	0.60	**1.00**	**1.00**	**1.00**	0.60	0.60	0.40

表 9：IntentBench 上编辑区域与非编辑区域的 F1 分数。粗体表示每列的最佳分数。

模型	编辑区域 F1 分数	未编辑区域 F1 分数
DocIntentOCR-7B	0.8624	0.8170
DocIntentOCR-3B	0.8624	**0.8321**
Qwen2.5-VL-7B-Instruct	0.7151	0.7338
Qwen2.5-VL-3B-Instruct	0.6699	0.7074
GPT-4.1	0.8204	0.6815
GPT-4o	**0.8677**	0.8206
o3	0.8613	0.8052
Claude Sonnet 4.6	0.8280	0.6888
Gemini 2.5 Flash	0.8140	0.7145
Qwen3-VL-32B-Instruct	0.7774	0.6041
Gemini 2.5 Pro	0.7710	0.6938
Llama-3.2-90B-Vision	0.6516	0.5601
Mistral-Large-3-675B	0.7204	0.6828

议如 IntentBench）。表10 报告了 F1 分数和 WER。

有两点观察。第一，即使面对域外文档，DocIntentOCR-3B 和 DocIntentOCR-7B 仍保持着相较于零样本视觉语言模型的性能优势，这表明意图感知监督能够泛化到标注解读任务，而非过拟合于金融特定内容。第二，各模型的相对排名在不同领域间基本保持不变，这说明 IntentBench 的排名是相当稳定的信号，而非领域伪象。我们强调，这只是一项初步试点（30 份文档，并非完整的基准测试集）；作为未来工作，我们承诺发布专门的跨领域基准测试，涵盖范围更广的文档。

表 10：跨领域试点研究（30 份文档：法律、医疗保健、房地产）。粗体表示各列中的最高分。

模型	F1	WER
GPT-4.1	0.8006	0.4677
Mistral-Large-3-675B	0.7698	0.4578
Gemini 2.5 Flash	0.8180	0.3580
Gemini 2.5 Pro	0.8520	0.2980
Llama-3.2-90B-Vision	0.8350	0.3270
Qwen3-VL-32B-Instruct	0.7502	0.6229
GPT-4o	0.8661	0.2343
o3	0.8749	0.2240
Qwen2.5-VL-7B-Instruct（基础版）	0.7971	0.4397
Qwen2.5-VL-3B-Instruct（基础版）	0.7664	0.4650
DocIntentOCR-3B	0.8961	0.1540
DocIntentOCR-7B	**0.9389**	**0.0873**

H DocIntentOCR-7B 与 GPT-4.1 的输出对比

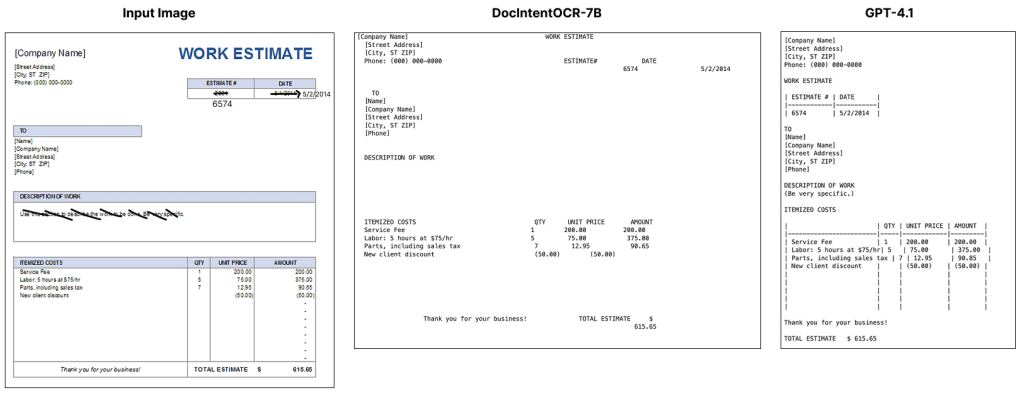


图 3：DocIntentOCR-7B 和 GPT-4.1 在带注释的金融文档上的输出比较。