

DrishtiKon: 面向文本丰富文档图像的多粒度视觉定位

Badri Vishal Kasuba, Parag Chaudhuri, Ganesh Ramakrishnan,

{badrivishalk, paragc, ganesh}@cse.iitb.ac.in

Department of Computer Science and Engineering

IIT Bombay, India

Abstract

文本丰富的文档图像中的视觉定位是文件智能和视觉问答 (VQA) 系统中的一个关键但未充分探索的挑战。我们提出了德里希特康, 这是一种多粒度的视觉定位框架, 旨在增强复杂、多语言文档中 VQA 的可解释性和可信度。我们的方法集成了健壮的多语言 OCR、大型语言模型和一种新颖的区域匹配算法, 以准确地在块级、行级、词级和点级定位答案范围。我们从 CircularsVQA 测试集中整理了一个新的基准数据集, 提供跨多个粒度的细粒度的人工验证注释。广泛的实验表明, 我们的方法达到了最先进的定位准确性, 行级粒度提供了精度与召回率之间的最佳平衡。消融研究进一步突出了多块和多行推理的好处。与领先的视觉语言模型的比较评估揭示了当前 VLM 在精确定位方面的局限性, 强调了我们基于结构化对齐的方法的有效性。我们的发现为更强大且可解释的真实世界、以文本为中心场景中的文档理解系统铺平了道路。代码和数据集已发布于 <https://github.com/kasuba-badri-vishal/DhrishtiKon>。

1. 介绍

视觉定位是在给定自然语言文本查询/问题作为输入的情况下, 在图像中定位与之对应的特定视觉元素的任务, 以生成一个可能的答案来回答来自输入图像 [8] 的查询。这项任务在视觉问答 (VQA) 任务的背景下尤其相关, 其目标是基于图像的视觉内容提供准确且上下文相关的答案, 并在预测答案的相关性上增加透明度和信任。此外, 针对该任务的地基学习允许更精

确和上下文相关的响应, 因为它要求模型直接关注与问题相关的视觉内容以找到地基答案。在这项工作中, 我们将探索文本中心图像上的视觉定位任务, 并评估在一个由政府公告图像集合构建的多粒度地基基准上的性能。我们提出了一种多粒度视觉定位框架 DhrishtiKon, 旨在增强模型在多个粒度级别上将答案地基化到视觉元素中的能力, 支持多页内容并采用一种新颖的区域匹配算法来改进地基过程。

2. 文献研究

多模态大型语言模型 (MLLMs) 的发展提升了文档理解, 但在文本丰富的文档图像中的视觉定位仍然是一个未充分探索但对文档智能任务至关重要的挑战。最近的努力如 TGDdoc [7]、DOGE-Bench [9] 和 TRIG-Bench [3] 引入了基准和模型来解决这一差距。表 1 比较了这些基准在数据、方法论、支持和合成数据生成策略方面的差异。

TGDdoc [7] 通过在 99K 张 PowerPoint 幻灯片和 12K 个 GPT-4 生成的对话上进行指令调优, 使用 BLIP2、PaddleOCR 和 GPT-4 来对齐文本识别与空间线索, 从而提高 MLLMs 的空间意识。

DOGE-Bench [9] 引入了图表、海报和 PDF 中的多粒度锚定和引用任务, 通过合成管道 (DOGE-Engine) 生成 70 万训练样本和一个 4K 评估基准, 实现了块级、行级和词级的细粒度引用。

相比之下, TRIG-Bench [3] 强调在真实世界、布局密集型文档中进行锚定, 使用混合 OCR-LLM-人类管道, 提供了 9 万合成样本和 800 个人工精选样本, 以评估 MLLMs 在空间复杂设置中的性能。

表 1. 文本丰富图像上的视觉定位基准比较

基准测试	图像来源	数量	粒度	支持	合成数据生成	
					光学字符识别	大语言模型
TGDoc	Online Powerpoint samples	99k 幻灯片 12k 高质量对话	Single	Single-Block	PaddleOCR	GPT4
DOGE-Bench	Charts, Posters, PDFs	700 千指令 4 千基准测试	Multi-Granular	Single-Block	易 OCR	GPT4o
TRIG-Bench	DocVQA, InfographicVQA	90k 合成 800 手动	Single	Multi-Block	PaddleOCR	GPT4o

总体而言，这些基准揭示了当前 MLLMs 处理密集布局的局限性，并强调了改进文档中心任务中空间推理的必要性。

虽然这些基准在视觉锚定方面取得了显著进展，但也暴露了当前 MLLMs 状态下的关键差距。

主要差距特别在于多语言、多块场景中的人类验证标注，适用于所有类型的富文本图像。

在这篇论文中，我们的目标是开发一个多语言、多块和多粒度的视觉锚定基准，以解决这些局限性，并为更强大且多功能的文档理解系统铺平道路。

3. DhrishtiKon 方法论

本节详细介绍了针对文本丰富的文档图像的多粒度视觉定位解决方案德里希特康 的建议流程。我们的方法集成了多语言 OCR [4,5]、大型语言模型 (LLMs) [2] 以及如第 3.4 节讨论的基于评分的混合区域匹配算法，以实现语义相关文本块的精确定位。图 1 提供了德里希特康 流程的概述。

3.1. 输入获取

我们从一张文本丰富的文档的输入图像开始，通常是扫描的官方通告或备忘录，这是问答的视觉基础。与图像相伴的是一个自然语言问题，询问文档中任何可提取的信息，如特定数据、引用或上下文细节。该问题是开放式的，允许一系列可能的答案，这些答案可能会跨越文档内的多个文本块，但其本质是提取性的，旨在从图像内容中找到具体的短语或数据点。

3.2. 多语言块级 OCR

输入图像使用 DocTR [4] 和 Surya-OCR [5] 的健壮多语言 OCR 系统进行处理，该系统通过布局预测将文

档分割成块级文本区域。每个区域都用其对应的边界框坐标和转录内容进行编码。这使得能够对文档进行空间分解以实现细粒度的语义定位。

3.3. 条件问题回答预测

提取的块，以及用户输入的提问，被输入到一个大型语言模型 (LLM) 中。我们使用 LLaMA-3.1-8B instruct 开源模型 [2] 作为我们的模型，该模型对文本内容进行语义推理并输出预测答案。这一步骤模拟了纯文本模式下的 MLLM (多模态大型语言模型) 推理，绕过了在答案生成过程中对视觉输入的需求，因为它将 OCR 输出作为上下文信息。

3.4. 候选区域匹配算法

为了在文档中定位预测的答案，我们实现了一个匹配算法，该算法遍历 OCR 引擎提取出的每个可能的文本块。对于每个区块，使用以下组件计算一个复合匹配分数：

- 模糊分数：**使用部分匹配和基于令牌的相似性匹配预测答案与块内容之间的相似度进行计算。
- 长度因子：**奖励与预测答案相对长度合理的区块，并通过可能匹配的文本进行比较。
- 惩罚函数：**降级那些具有过短边界框（表明是噪声）或缺乏上下文语义重叠的区块。

如算法 1 所述，每个候选区域根据其组成部分进行评分，这些部分又具有不同的贡献比例，得分超过预定义阈值的区域将被保留。候选块根据它们的综合分数进行排序。选择并可视化前 k 个匹配的块以及预测答

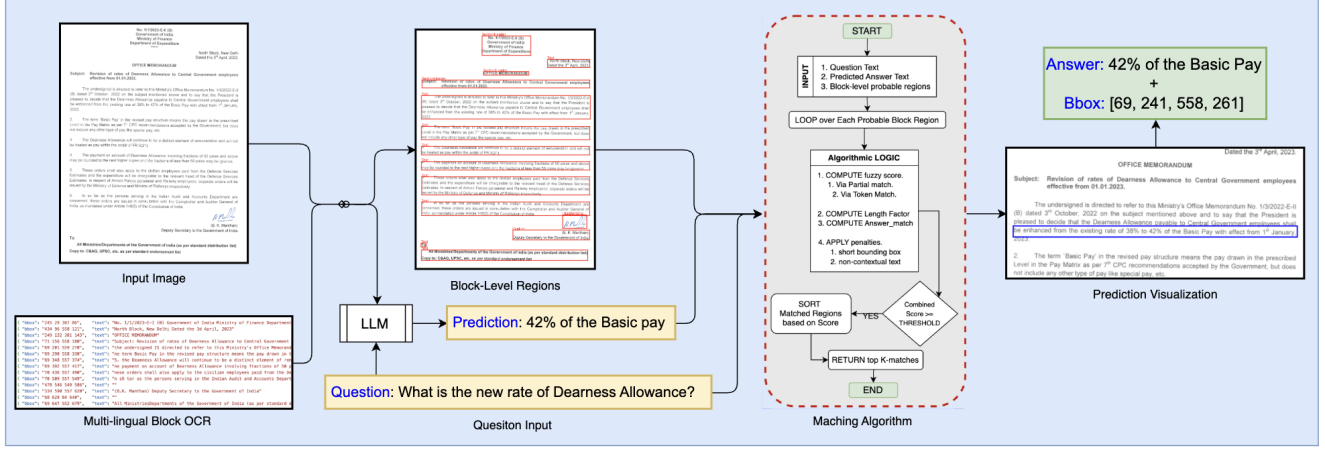


图 1. 德里希特康：多粒度视觉定位解决方案流水线，在示例说明中包含匹配算法流程图的 VQA 样本。

案以提高可解释性。这允许基于阈值参数选择所需的多块预测数量。

Algorithm 1 块级潜在区域的区域匹配算法

- 1: **输入:**
- 2: 1. 问题文本 Q
- 3: 2. 预测答案文本 A
- 4: 3. 块级区域 $R = \{r_1, r_2, \dots, r_n\}$
- 5: for each region $r \in R$ do
- 6: **步骤 1: 计算模糊分数**
- 7: a. 通过部分匹配 A 和 r 中的文本
- 8: b. 通过标记匹配 A 和 r 中的文本
- 9: **步骤 2: 计算长度因子**
- 10: 根据边界框/文本长度计算比率或惩罚
- 11: **步骤 3: 计算答案匹配得分**
- 12: 评估 A 和 r 内容之间的相似性
- 13: **步骤 4: 应用惩罚**
- 14: a. 对短边界框进行惩罚
- 15: b. 对非上下文或无关文本进行惩罚
- 16: **步骤 5: 计算综合评分**
- 17: if Combined Score $\geq$ Threshold then
- 18: 将 r 标记为有效匹配
- 19: end if
- 20: end for
- 21: **步骤 6: 排序** 基于综合得分匹配的区域
- 22: **步骤 7: 返回** 前 K 个匹配结果

4. 实验与结果

4.1. 基准数据集

为了评估我们的视觉定位框架的有效性，我们整理了一个基准数据集，该数据集源自印度政府公告。我们采样了 70 种多样的文档图像，涵盖了复杂的布局结构和多语言文本内容，使其适合于评估细粒度的空间语义推理。该数据集包括如表 2 中总结的详细注释。正如第 2 节所述，现有的基于文本中心图像的视觉定位数据集 TGDdoc [7]、DOGE-Bench [9] 和 TRIG-Bench [3] 尚未发布且不可用，因此我们在整理好的测试集中进行了实验测试，该测试集包含带有定位的 509 个问答对。

表 2. 来自 CircularsVQA 测试集的标注数据集的多粒度视觉定位统计。

属性	计数
Document Images	70
Question – Answer (QnA) Pairs	509
Block-level regions	538
Line-level regions	988
Word-level regions	5,968
Point-level data	538

4.1.1 注释工具

为了构建我们在实验中使用的标注测试集，我们开发了一个内部定制的注释工具，旨在支持文档区域

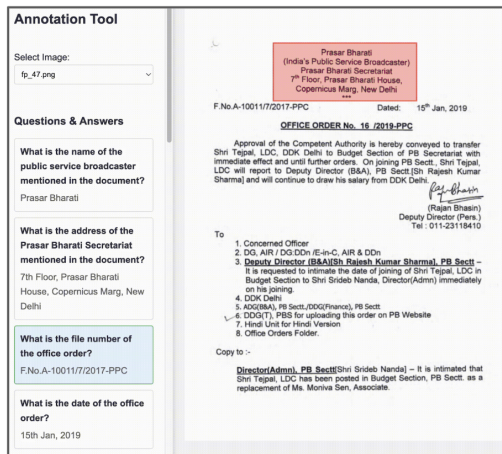


图 2. 内部标注工具支持多级区域标记。

的层次化和细粒度注释。该工具在**块、行、单词和点**级别上进行标注，提供了定义与自然视觉布局匹配的答案跨度的灵活性。该工具具有一个交互式界面（图 2），允许标记精确的跨度和关系，并且还支持从行级别注释区域自动检测单词级别的文本，有助于增强注释。与其他现有数据集 [7, 9] 不同，本数据集是通过人工标注完成的，而不是使用专有的大语言模型生成合成数据。

4.2. 按粒度评估

我们使用标准指标评估所提出的视觉定位管道的性能：**精度、回忆和 F1 分数**。这些指标在四个级别的定位粒度上进行计算：块、线、单词和点。结果，如表 3 所示，对应于使用我们的区域匹配算法在问题 + 基准答案输入下生成的预测。图 3 说明了这些粒度级别的视觉定位定性结果。

表 3. 不同粒度的接地准确性评估

级别	精度 (%)	召回率 (%)	F1 分数 (%)
Block	53.89	68.22	60.21
Line	65.06	73.68	69.10
Word	60.32	49.87	54.60
Point	60.66	51.30	55.79

我们观察到**逐行接地**达到了最高的 F1 分数 69.10%，这表明这种粒度在精度和召回率之间提供了最佳的平衡。这可能是由于政府公告的结构性质，其中答案通常跨越单个格式良好的行。**块级接地**从捕获语

义完整的单元中受益，但遭受区域碎片化和不一致的边界框聚合的影响。同时，尽管**词级和点级别**定位更为细粒度，但也更容易受到 OCR 错位引起的错误的影响，这导致了召回率降低。总体而言，这些发现表明中间级别的粒度（尤其是行级别）对于结构化文档中的视觉定位是最优的。然而，正如最近的工作 MOLMO [1] 所强调的那样，点级微调仍然对训练有益，因为它提供了所需的最小定位上下文并减少了培训成本。

4.3. 按模型评估

我们进一步将区域匹配算法与多个视觉语言模型 (VLMs) 进行了比较，包括 LLaMA 3.1 [2] 和 Qwen2.5-VL [6]。所有模型都在不同监督方案下的行级定位任务中进行了评估：使用问题和预测答案（问答），仅使用问题，以及结合预测答案与定位算法的混合配置。结果如表 4 所示。

区域匹配算法与真实答案配对时，明显优于所有其他方法，并且可以被认为是该数据集中行级定位任务的理想条件下的顶点性能。

Qwen2.5-VL [6] 虽然是一个通用的多模态模型，在所有指标上的表现都非常差。这表明当前无 OCR 和 VLM 方法在处理以文本为中心的视觉定位任务时存在局限性。

值得注意的是，在提供的 OCR 输入上下文中，LLaMA 3.1 在预测视觉接地方面表现出竞争力，表明当给定结构化的文本上下文时，大语言模型可以有效地推理空间关系。

我们还在表 5 中进行了块级别的比较评估。区域匹配算法在块级别上再次表现出优越性能，强调了其在空间推理方面的一贯优势。性能差距也突显了布局感知对齐在真实世界文档问答设置中的重要性。

4.4. 消融研究

为了理解在所提出的区域匹配算法中支持多块和多行区域的影响，我们在推理过程中对允许的块数和行数进行了详细的消融研究以提高准确性。

最大块的影响：

我们在区域匹配算法中评估了性能，通过将块的最大数量从 1 变到 5，并使用问题+地面实况答案对齐。如表 6 所示，允许多个块会提高召回率但以牺牲精确率为代价。当允许最多 2 个块时，获得了最佳的 F1 得分

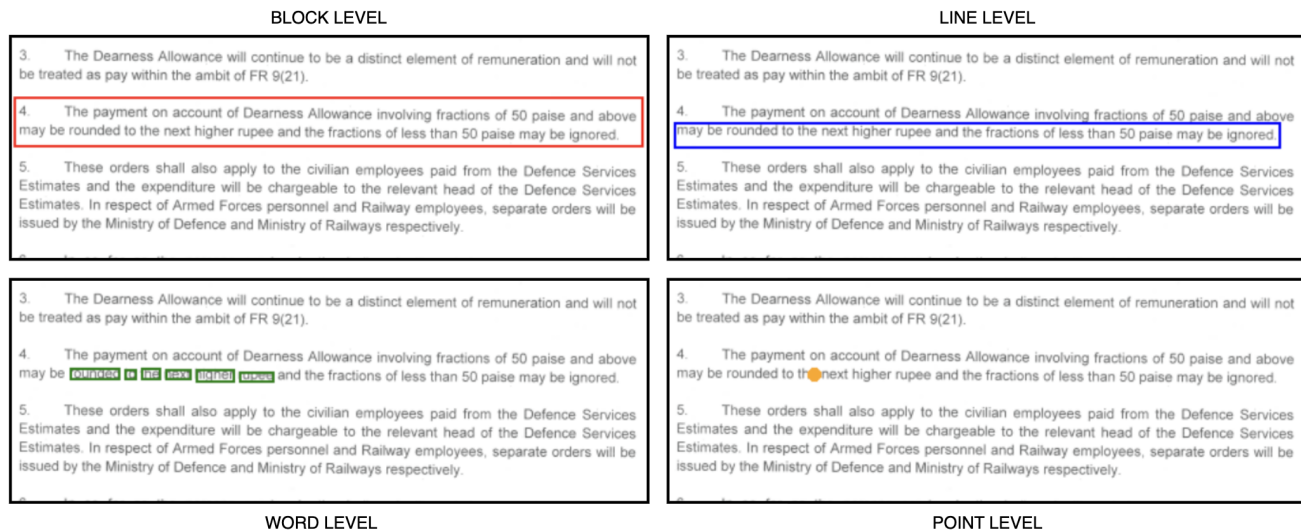


图 3. 不同粒度的视觉定位结果：块、行、词和点。边界框表示与问题和答案对应的预测区域。

表 4. 逐行接地评估（最多 10 个框，IoU=0.5），基于 OCR 的输入（文本+边界框）。

输入（文本 + 边界框）	光学字符识别（问答）	模型（问答）	OCR（地面）	模型（地面）	P (%)	R (%)	F (%)
Ground Truth Answer	-	-	YES	Algorithm	65.06	73.68	69.10
Ground Truth Answer	-	-	YES	LLAMA	49.20	64.98	56.00
Ground Truth Answer	-	-	No	Qwen2.5VL	6.00	4.55	5.18
Predicted Answer	YES	LLaMA	YES	Algorithm	43.97	53.14	48.12
Predicted Answer	YES	LLaMA	YES	LLAMA	37.43	47.77	41.97
Predicted Answer	NO	PATRAM	YES	Algorithm	35.58	30.97	33.12

(62.68%)，这表明确实存在多块答案，但是超过两个块的过度聚合会引入噪声。

最大行的影响：

我们进一步扩展了对答案跨度中允许的行数进行的消融研究。表 7 和图 4 显示了在迭代过程中准确率、召回率和 F1 分数的变化。随着聚合的行数增加，准确率稳步下降，而召回率则持续提高，在大约 10 行时达到平稳期。F1 分数稳定在 69.3% 左右，在 5 行时达到了最优性能。

本研究强调了一个关键的权衡：精度在较低聚合水平上占主导地位，而随着答案跨度的增长，召回率则得到改善。允许多行、多块推理显著提升了性能，但必须谨慎限定以避免降低精度。对于我们的任务，2 块和 5 行的聚合提供了最均衡的性能。

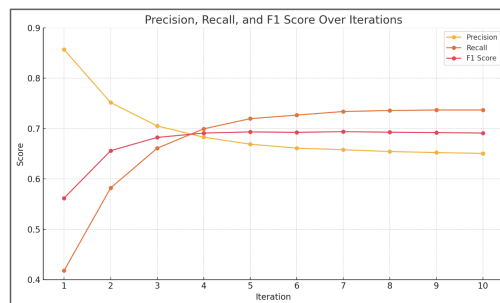


图 4. 精确率、召回率和 F1 分数随最大行阈值增加的趋势。

5. 结论与未来工作

在这项工作中，我们提出了德利什提康，一个旨在提高基于文本的图像上 VQA 系统可解释性和可信度的多粒度视觉定位框架。通过整合多语言 OCR、大型语言模型和一种新颖的区域匹配算法，我们在多个粒度

表 5. 块级接地评估跨方法

方法	精度 (%)	召回率 (%)	F1 分数 (%)
Region Matching Algorithm	53.89	68.22	60.21
LLaMA 3.1 (Q+A)	41.72	67.47	51.56
LLaMA 3.1 (Question Only)	32.88	54.46	41.01

表 6. 最大块数对区域匹配性能的影响。

最大块数	精度	回忆	F1 分数
1	71.83	61.15	66.06
2	58.67	67.29	62.68
3	55.57	67.66	61.02
4	54.53	68.22	60.61
5	53.89	68.22	60.21

表 7. 最大行数对区域匹配性能的影响。

最大行数	精度	回忆	F1 分数
1	85.68	41.80	56.19
2	75.16	58.20	65.60
3	70.52	66.09	68.23
4	68.28	69.94	69.10
5	66.89	71.96	69.33
6	66.11	72.67	69.24
7	65.79	73.38	69.38
8	65.44	73.58	69.27
9	65.23	73.68	69.20
10	65.06	73.68	69.10

级别上显著提高了定位准确性。在从 CircularsVQA 测试集提取的一个精选数据集上的实验表明，中等粒度（行级）对于基于文档的问题回答任务表现最佳。消融研究进一步验证了多块和多行推理在提高定位精度和召回率方面的有效性。此外，评估结果还显示，尽管大型视觉语言模型可以产生合理的语义预测，但在没有结构化的定位机制的情况下，它们在精确的视觉定位任务上仍显不足。我们的区域匹配方法证明，当结合稳健的 OCR 和文本匹配时，传统的基于对齐的方法仍然对于需要空间保真的文档理解任务具有高度竞争力。

参考文献

[1] Matt Deitke, Christopher Clark, Sangho Lee, and team. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. 4

[2] Abhimanyu Dubey and team. The llama 3 herd of models. ArXiv, abs/2407.21783, 2024. 2, 4

[3] Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Deroncourt, Wanrong Zhu, Tianyi Zhou, and Tongfei Sun. Towards visual text grounding of multimodal large language model. ArXiv, abs/2504.04974, 2025. 1, 3

[4] Mindee. doctr: Document text recognition. 2021. 2

[5] Vikas Paruchuri and Datalab Team. Surya: A lightweight document ocr and analysis toolkit. <https://github.com/VikParuchuri/surya>, 2025. GitHub repository. 2

[6] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024. 4

[7] Yonghui Wang, Wen gang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms. ArXiv, abs/2311.13194, 2023. 1, 3, 4

[8] Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. ArXiv, abs/2412.20206, 2024. 1

[9] Yinan Zhou, Yuxin Chen, Haokun Lin, Shuyu Yang, Li Zhu, Zhongang Qi, Chen Ma, and Ying Shan. Doge: Towards versatile visual document grounding and referring. ArXiv, abs/2411.17125, 2024. 1, 3, 4