Nayana: A Foundation for Document-Centric Vision-Language Models via Multi-Task, Multimodal, and Multilingual Data Synthesis

Adithya S Kolavi
CognitiveLab
Bengaluru, India
adithyaskolavi@gmail.com

Samarth P
CognitiveLab
Bengaluru, India
samarthprakash8@gmail.com

Vyoman Jain
CognitiveLab
Bengaluru, India
vyomanjain@gmail.com

Abstract

We present Nayana, a comprehensive, synthetically generated dataset designed to advance document-centric vision language models across multiple tasks and languages. Nayana consists of three interconnected subsets, each targeting different aspects of document understanding: (1) a base dataset of 3 million document images with hierarchical annotations including detailed layout information, textual content, reading order, and relationships between document elements; (2) a multilingual variant spanning 22 languages, preserving the original document layout while translating text through contextual models; and (3) a specialized information retrieval subset for document ranking tasks with approximately 250,000 image-query pairs per language.

What distinguishes Nayana is its synthetic generation methodology. We collect a diverse corpus of PDFs from multiple sources, then apply state-of-the-art models to hierarchically extract structural and textual information, yielding a highly structured representation capturing layout elements, text lines, images, captions, and their interrelationships. For the multilingual extension, we employ contextual translation models to transform textual elements while preserving stylistic and visual attributes. Beyond the primary subsets, Nayana incorporates Visual Question Answering (VQA) pairs in both monolingual and multilingual settings.

This multifaceted approach makes Nayana a truly multi-task dataset, enabling training of vision-language models for diverse applications including layout detection, equation recognition, image captioning, markdown conversion, multilingual OCR, document retrieval, and more.

1. Introduction

Document understanding remains a fundamental challenge at the intersection of computer vision and natural language processing. As digital document collections grow exponentially across domains, intelligent systems capable of processing, analyzing, and retrieving information from visually rich documents have become increasingly vital. Although significant progress has been made in recent years with the advent of vision-language models (VLMs), these advances have been constrained by the scarcity of large-scale, diverse, and comprehensively annotated document datasets that capture the complex interplay between textual content, visual elements, and document layout.

The limitations of existing document datasets are multifaceted. Many are domain-specific (e.g., scientific papers, business forms), limited in scale, or lack hierarchical annotations that preserve the structural relationships between document elements. Furthermore, multilingual document understanding remains underexplored, despite its critical importance in our global information ecosystem.

To address these challenges, we present Nayana (meaning "eyes" in Hindi), a comprehensive, synthetically generated dataset designed to advance document-centric vision-language models across multiple tasks and languages. Nayana consists of three interconnected subsets, each targeting different aspects of document understanding:

1. **Base Corpus:** A collection of 3 million document images with hierarchical annotations, including detailed layout information, textual content, reading order, and relationships between document elements. This subset provides rich structural representations at multiple levels of granularity, which is essential for document understanding tasks.
2. **Multilingual Dataset:** An extension of the base corpus comprising 2.2 million document images in 22 languages. Unlike existing multilingual datasets, Nayana preserves the original document layout while translating textual content through contextual translation models, ensuring stylistic consistency across languages.
3. **Retrieval Dataset:** A specialized subset designed for document ranking tasks, containing approximately 250K image-query pairs per language. This enables the training of multimodal retrieval models to effectively rank document images based on textual queries.

The significance of Nayana lies in its comprehensive approach to document understanding. Each document image in the dataset is accompanied by annotations for multiple tasks, allowing researchers to train and evaluate models across different document understanding challenges using a consistent data source. We demonstrate Nayana's effectiveness as a multi-task training resource by showing substantial improvements in OCR accuracy, document retrieval, and multilingual document processing compared to existing datasets and models.

Additionally, we present a novel layout-preserving translation pipeline that generates 2.2 million multilingual document images across 22 languages while maintaining visual and structural consistency, significantly advancing multilingual document understanding capabilities.

2. Related work

2.1. Document Understanding with Vision-Language Models

Vision-language models have revolutionized document understanding by jointly modeling text, layout, and visual features. LayoutLM [23] pioneered this approach by pre-training on text and layout information for document image understanding, achieving state-of-the-art results in form and receipt understanding tasks. Building on this foundation, LayoutLMv2 [24] incorporated image features to enhance performance on tasks requiring visual context. Several other models like Donut [13] and Nougat [5] have also been developed for visual document understanding.

Recent advances include DocLLM [20], a lightweight extension to large language models that focuses on layout-aware processing using bounding box information. This approach avoids expensive image encoders while maintaining efficiency, making it suitable for resource-constrained environments.

2.2. Synthetic Data Generation

Synthetic data generation has emerged as a cost-effective solution for training machine learning models, particularly when real-world data is scarce or raises privacy concerns. Tools like genalog [15] generate document images that mimic scanned analog documents while preserving layout and structure, providing valuable training data for OCR and layout analysis tasks.

Synthetic Document Generator for Annotation-free Layout Recognition [16] proposed a process for generating realistic documents by modeling the document's components as random variables using Bayesian networks. More recently, DocSynth-300K [26] introduced a large-scale synthetic dataset for document layout analysis, demonstrating the value of diverse synthetic data in enhancing model accuracy and generalization capabilities.

2.3. Multimodal and Multilingual Document Processing

Addressing the challenges of linguistic diversity in document processing requires models capable of handling multiple languages and modalities simultaneously. LayoutXLM [11] extends multimodal pre-training to multilingual scenarios, bridging language barriers in visually-rich document understanding.

Pangea [25] introduced PangeaInstruct, a 6M multilingual multicultural multimodal instruction tuning dataset spanning 39 languages. These approaches highlight the growing need for models that can process documents in their full multimodal and multilingual complexity.

2.4. Document Ranking for Information Retrieval

Document ranking is a critical component of information retrieval systems, particularly in multimodal contexts. ColPali [9] is an efficient document retrieval method using vision-language models that directly embeds page images for retrieval and ranking, demonstrating superior performance on the ViDoRe [9] benchmark.

Our work builds upon these foundations, addressing gaps in current research particularly related to large-scale multilingual document understanding with preserved layout information and comprehensive hierarchical annotations.

3. The Nayana Dataset: Overview and Methodology

Nayana is a large-scale, synthetically generated dataset designed for document-centric VLM advancement. Its creation was guided by these key principles: hierarchical representation (from page-level to lines, figures, equations), layout preservation (crucial for visual integrity), multilingual support (maintaining structural consistency across languages), multi-task utility (annotations supporting OCR, layout analysis, VQA, retrieval, etc.), and scalable, privacy-preserving synthetic generation.

3.1. Data Collection and Composition

We curated approximately 3 million document images from diverse public sources like arXiv, PubMed, SafeDocs [1], Docmatix [12], government archives, and technical manuals, ensuring broad domain coverage (scientific literature 55%, government/technical docs 20%, others 25%).

3.2. Hierarchical Extraction Pipeline

Nayana's core methodology transforms raw PDFs into richly annotated structures using a multi-stage pipeline leveraging ensemble approaches for robustness:

1. Rendering & Metadata: PDFs are rendered at 300 DPI. Baseline metadata (text, fonts) is extracted via PyMuPDF [2] as a complementary source.

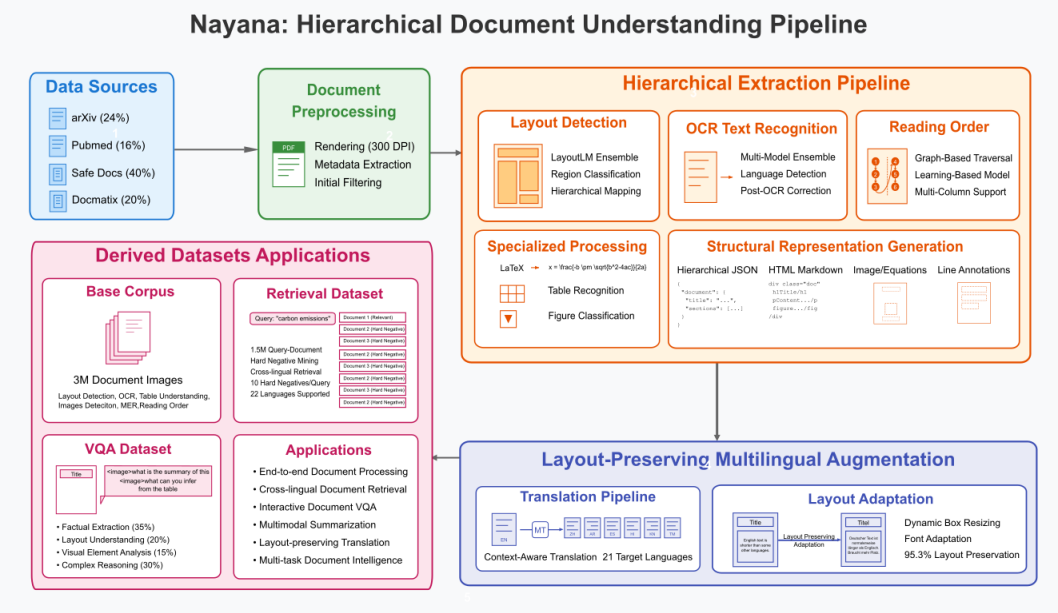

Figure 1. Nayana's Hierarchical Extraction Pipeline

Table 1. Distribution of Document Sources in Nayana

Source	Number of Document Images	Percentage
arXiv	600,000	20.0%
PubMed	400,000	13.3%
Safe Docs	1,000,000	33.3%
Docmatix	565,000	18.8%
Others	435,000	14.5%

2. Layout Detection: An ensemble (incl. LayoutLM [23], DocLayout-YOLO [27]) performs hierarchical detection (titles, text blocks, tables, figures, etc.). A consolidation algorithm resolves conflicts while preserving structure.
3. OCR & Text Recognition: We integrate Tesseract [17] and PaddleOCR [8]) to extract text while balancing speed and accuracy for complex cases (e.g. scientific notation, complex layouts).
4. Reading Order Detection: A hybrid approach (rule-based heuristics + learning models [21, 22]) determines logical content flow.
5. Specialized Element Processing: Mathematical expressions are identified and converted to LaTeX via Unimernet [19], and the figures are extracted and classified.
6. Structural Representation: Outputs include hierarchical JSON, Markdown, HTML (preserving visual styling), and fine-grained line-level annotations, supporting diverse model inputs and multi-objective training.

3.3. Key Dataset Components Generation

Layout-Preserving Multilingual Augmentation: Our layout-preserving augmentation pipeline transforms the base corpus into multilingual versions (22 languages spanning diverse families/scripts). A hybrid translation engine combines Google Translate API and specialized language models fine-tuned [10] for document translation is applied to text segments. Crucially, the fonts are randomized and layout adaptation dynamically adjusts bounding boxes to maintain visual and structural consistency across languages, preserving the original document's look and feel.

Visual Question Answering (VQA) Dataset: Leverages the rich annotations to generate diverse QA pairs using advanced LLMs/VLMs (e.g., DeepSeek-V3 [7], PaliGemma [4], Qwen2.5 VL [3]) for complex, multimodal reasoning (factual, layout, visual). Pairs are generated for the base language and translated to create parallel monolingual and cross-lingual VQA examples.

Document Retrieval Dataset: Addresses the need for document ranking models. Queries are derived from VQA and a hard negative mining strategy is employed: documents are embedded using multiple multimodal models (Jina CLIP v2 [14], Jina Embeddings v3 [18], BGE-M3 [6]), and semantically similar but non-ground-truth documents are identified and ranked as hard negatives. This contrastive learning setup (25 hard negatives per query) enables models to learn fine-grained distinctions across 22 languages.

This multi-faceted approach makes Nayana a comprehensive resource for advancing document intelligence across tasks and languages, particularly enabling research in layout-aware multilingual understanding and robust document retrieval.

4. Results and Evaluation

We evaluated the performance of models trained on Nayana across multiple document understanding tasks, including OCR, document VQA and document retrieval. Our experiments demonstrate significant improvements using a variety of finetuned models on subsets of the Nayana dataset such as Nayana-OCR (a GOT OCR 2.0 finetune), a finetuned Qwen 2.5 VL 3B and document retrieval models like Colpali compared to the respective baseline approaches.

4.1. Document Visual Question Answering

Table 2 presents the results on document visual question answering benchmarks. The metrics include accuracy on DocVQA, InfographicVQA, and MMMU datasets.

Table 2. Document Visual Question Answering Performance (Accuracy, higher is better)

Model	**DocVQA**$_{test}$	**InfographicVQA**$_{test}$	**MMMU**$_{val}$
InternVL2.5-8B	93.0	77.6	56.0
Qwen2-VL-7B	94.5	76.5	54.1
Qwen2.5-VL-7B	**95.7**	**82.6**	**58.6**
Qwen2.5-VL-3B Finetune	95.7	82.3	58.4

4.2. OCR and Text Extraction Performance

Table 3 compares the performance of various vision-language models on text extraction tasks such as extracting markdown or HTML from images. The metrics include Edit Distance, F1-score, Precision, Recall, BLEU, and METEOR scores. The results highlight the superiority of the model fine-tuned on Nayana, which achieves the best precision and competitive performance across all metrics.

Table 3. Performance comparison of different vision-language models on text extraction tasks in English

Model	Edit Distance ↓	F1-score ↑	BLEU ↑	METEOR ↑
LLaVA-NeXT (34B)	0.430	0.647	0.478	0.582
InternVL-ChatV1.5 (26B)	0.393	0.751	0.568	0.663
Nougat (250M)	0.255	0.745	0.665	0.761
Vary (7B)	0.092	0.918	0.885	0.926
Vary-toy (1.8B)	0.082	0.924	0.889	0.929
Qwen-2VL (7B)	0.096	0.931	0.893	0.936
Qwen-2.5-VL Finetune (3B)	**0.046**	**0.952**	**0.930**	**0.954**

Table 4 presents the average OCR performance metrics across ten Indic languages evaluated using a new benchmark, Nayana OCR bench comprising 160 samples of multilingual images and the corresponding text.

4.3. Document Retrieval Performance

Table 5 compares the document retrieval performance of the baseline models with those fine-tuned in subsets of the Nayana retrieval dataset. For evaluation, we introduce Nayana-IR DescVQA, an evaluation benchmark of 300 curated document-query pairs across 10 languages.

Table 4. Average performance metrics across all evaluated languages. (Refer Table **??** for per-language results)

Model	**CER↓**	**WER↓**	**BLEU↑**	**ANLS↑**	**METEOR↑**
Tesseract	**0.206**	0.583	0.318	**0.797**	0.540
PaddleOCR	0.621	0.880	0.020	0.287	0.069
Llama-3.2 11B	3.858	3.900	0.007	0.091	0.055
Phi-3.5 Vision	2.420	2.461	0.007	0.086	0.044
Qwen2-VL 2B	1.776	1.793	0.025	0.129	0.086
GOT-OCR	0.945	1.041	0.016	0.071	0.052
Nayana-OCR	0.227	**0.463**	**0.395**	**0.796**	**0.630**

Table 5. Document retrieval performance comparison (For more detailed comparisons, refer to Table **??**)

Model	**Language**	**NDCG@5↑**	**NDCG@10↑**	**Recall@1↑**
ColPali v1.3	Hindi	0.513	0.531	0.432
	Kannada	0.280	0.300	0.213
ColSmol 500M	Hindi	0.109	0.124	0.074
	Kannada	0.038	0.046	0.020
Nayana IR Finetuned Models				
ColPali v1.3 (Hi and Kn Finetune)	Hindi	0.651	0.673	0.550
	Kannada	**0.493**	**0.518**	**0.383**
ColSmol Base (Hindi Finetune)	Hindi	**0.729**	**0.741**	**0.661**
	Kannada	0.282	0.311	0.19

4.4. Future Work and Conclusion

Our experiments demonstrate the significant impact of Nayana, our comprehensive synthetic dataset for document intelligence. Models finetuned on Nayana show marked performance improvements across multiple tasks. Notably, Nayana-OCR surpasses established systems like Tesseract and PaddleOCR, while our Nayana-IR finetuned models, such as ColPali v1.3, yield substantial gains in multilingual document retrieval highlighting the value of high-quality, task-specific data provided by Nayana.

While Nayana represents a significant advance, future work includes expanding its diverse but incomplete language coverage (currently 22 languages), particularly to low-resource languages, facilitated by our modular translation pipeline. Furthermore, leveraging the existing rich dataset, we plan extensive experimentation involving further fine-tuning across various models and architectures. Our goal is to establish new state-of-the-art results across the diverse spectrum of document understanding tasks enabled by Nayana's comprehensive annotations. In conclusion, Nayana addresses critical gaps through its scale, hierarchical structure, layout-preserving multilingual capabilities, and multi-task design, supporting unified training for real-world complexity. By providing this scalable, privacy-conscious, and extensible resource to the community, we aim to accelerate progress towards truly comprehensive multimodal and multilingual document understanding systems.

References

[1] SAFEDOCS Corpus. Retrieved from digitalcorpora.org, 2023. 2

[2] Inc. Artifex Software and contributors. PyMuPDF. GitHub repository, 2022. Python bindings for the MuPDF library. Available at https://github.com/pymupdf/PyMuPDF. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel M. Salz, Maxim Neumann, Ibrahim M. Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Martin Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bovsnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiao-Qi Zhai. Paligemma: A versatile 3b vlm for transfer. *ArXiv*, abs/2407.07726, 2024. 3

[5] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents, 2023. 2

[6] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. 3

[7] DeepSeek-AI. Deepseek-v3 technical report, 2024. 3

[8] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020. 3

[9] Arnaud Faysse, Piotr Bojanowski, Armand Joulin, Matthieu Cord, and Matthijs Douze. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024. 2

[10] Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023. 3

[11] Yiheng Hong, Yang Xu, Lei Cui, Yiheng Wei, Zhenfei Wang, Binyuan Zhang, Wenhao Hu, Jindong Wang, Weihong Liu, Furu Meng, et al. Layoutxlm: Multimodal pre-trained model for multilingual visually-rich document understanding. *arXiv preprint arXiv:2310.03302*, 2023. 2

[12] Hugging Face. Docmatix Dataset. Hugging Face Hub, 2024. A dataset for Document Visual Question Answering, featuring 2.4 million images and 9.5 million Q/A pairs. Available at https://huggingface.co/datasets/HuggingFaceM4/Docmatix. 2

[13] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, pages 322–339. Springer, 2022. 2

[14] Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images, 2024. 3

[15] Microsoft. genalog: Synthetic document generator. https://github.com/microsoft/genalog, 2021. 2

[16] Nishant Raman, Yannis Papanikolaou, Shinjiro Kanazawa, Rushin Vasani, Vishal Rathod, and Oluwasanmi Koyejo. Synthetic document generator for annotation-free layout recognition. *Pattern Recognition Letters*, 159:1–8, 2022. 2

[17] Ray Smith. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA, 2007. IEEE Computer Society. 3

[18] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task lora, 2024. 3

[19] Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. Unimernet: A universal network for real-world mathematical expression recognition, 2024. 3

[20] Jiasheng Wang, Jinghan Zhang, Chenhui Huo, Rui Wu, Zhicheng Liang, Xinyu Zhang, Chengzhe Sun, Nian Liu, Yongxin Zhang, Tao Guo, et al. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*, 2024. 2

[21] Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. ReadingBank Dataset, 2021. A benchmark dataset for reading order detection. Available at https://github.com/doc-analysis/ReadingBank. 3

[22] Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. LayoutReader: Pre-training of Text and Layout for Reading Order Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4746. Association for Computational Linguistics, 2021. 3

[23] Yang Xu, Minghao Li, Lei Cui, Shaohan Lu, Furu Ye, Hongming Ding, Zhoujun Huang, Gan Yang, Jinlan Zhou, Xinxing Wang, Jun Tang, Jianfeng Huang, Shan Yao, Jiawen Guan, Ji Zhang, Dayang Cai, Ming Li, Tat-Seng Chua, Eric Gao, Dacheng Tao, Trevor Cohn, and Shai Geva. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1192–1200, 2020. 2, 3

[24] Yang Xu, Yiheng Wei, Lei Cui, Wenhao Hu, Jinan Zeng, Minghao Li, and Furu Yao Lin Zhang. Layoutlmv2: Multimodal pre-training for visually-rich document understanding. In *International Conference on Learning Representations*, 2021. 2

[25] Xiang Yue, Yueqi Song, Akari Asai, Simran Khanuja, Anjali Kantharuban, Seungone Kim, Jean de Dieu Nyandwi, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*. 2

[26] Yechen Zhao, Taiming Zhang, Daniel Tam, Zeya Zeng, Haobo Yuan, Jindong Gu, Evan Liang, Dahua Lin, and Ji Lin. Docsynth-300k: A large-scale synthetic dataset for document layout analysis. *arXiv preprint arXiv:2410.12628*, 2024. 2

[27] Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception, 2024. 3