IndicDLP: A Foundational Dataset for Multi-Lingual and Multi-Domain Document Layout Parsing

Oikantik Nath[1][0009-0005-5407-0455], Sahithi Kukkala[2][0009-0003-5483-2232], Mitesh Khapra[1][0009-0008-3687-9922], and Ravi Kiran Sarvadevabhatla[2][0000-0003-4134-1154]

[1] Indian Institute of Technology, Madras, India
{oikantik,miteshk}@cse.iitm.ac.in
[2] International Institute of Information Technology Hyderabad, India
{ravi.kiran@,sahithi.kukkala@research.}iiit.ac.in

Abstract. Document layout analysis is essential for downstream tasks such as information retrieval, extraction, OCR, and digitization. However, existing large-scale datasets like PubLayNet and DocBank lack fine-grained region labels and multilingual diversity, making them insufficient for representing complex document layouts. Human-annotated datasets such as *M*^{6}*Doc* and D^{4}LA offer richer labels and greater domain diversity, but are too small to train robust models and lack adequate multilingual coverage. This gap is especially pronounced for Indic documents, which encompass diverse scripts yet remain underrepresented in current datasets, further limiting progress in this space. To address these shortcomings, we introduce *IndicDLP*, a large-scale foundational document layout dataset spanning 11 representative Indic languages alongside English and 12 common document domains. Additionally, we curate UED-mini, a dataset derived from DocLayNet and *M*^{6}*Doc*, to enhance pretraining and provide a solid foundation for Indic layout models. Our experiments demonstrate that fine-tuning existing English models on *IndicDLP* significantly boosts performance, validating its effectiveness. Moreover, models trained on *IndicDLP* generalize well beyond Indic layouts, making it a valuable resource for document digitization. This work bridges gaps in scale, diversity, and annotation granularity, driving inclusive and efficient document understanding.

Keywords: Document Layout Parsing · Indic Languages · Historical and Modern Documents

1 Introduction

Document Layout Parsing (DLP), also known as Document Layout Segmentation, is a fundamental task in document understanding that enables down-

Project Page: https://indicdlp.github.io/

stream applications such as information extraction [13], retrieval [16], OCR [12], and automated document conversion [5]. However, DLP remains highly challenging due to the diverse layouts, font styles, and structural variations found across different document types. Given its importance, several datasets have been introduced in recent years, either automatically derived from existing digital document metadata [32,18] or manually labeled [10,11,22]. While automatically generated datasets offer large-scale training data, they often lack fine-grained annotations. In contrast, manually labeled datasets provide higher-quality annotations, but are limited in scale. A common limitation across both types is the lack of multilingual coverage, with most datasets supporting only English and a few resource-rich languages. This resource gap has significantly hindered progress for low-resource languages, especially those from the Indian subcontinent, which feature diverse scripts distinct from Latin-based ones. As our results shall demonstrate (Figure 3), models trained on English documents have subpar generalization to Indian languages.

To bridge this gap, we introduce *IndicDLP*, the largest and most diverse Indic document layout dataset to date, consisting of 121,198 human-annotated images across 12 domains, 11 Indic languages alongside English, and 42 region labels. The dataset enables robust training for layout parsing across a wide range of document types, including newspapers, magazines, novels, textbooks, acts & rules, notices, manuals, syllabi, question papers, forms, brochures, and research papers. To ensure a comprehensive yet minimal label set, we aligned our annotation schema with the M^6Doc [10] guidelines, refining label granularity through manual analysis of *IndicDLP* images. The annotation process followed a maker-checker workflow with over 50 annotators and reviewers per language, guided by a detailed 150-page manual. Over 60% of the annotations were further validated for cross-domain consistency by a team of 8 supercheckers.

Using *IndicDLP*, we conduct a series of experiments to evaluate the performance of existing object detection and layout parsing models on Indic document layout parsing. We establish baseline performance by fine-tuning state-of-the-art models on *IndicDLP*, including YOLOv10 [27], DiT [17], DocLayout-YOLO [31], DINO [29], RoDLA [9], and a vision language model Florence-2 [28]. Our results show significant performance variations across domains, with models struggling on documents with more unique regions and complex multicolumn layouts. Performance also varies substantially across languages, with zero-shot evaluation on unseen languages showing a 20–25 mAP point drop compared to those seen during training, suggesting that script-based variations influence layout parsing performance.

Additionally, we explore the utility of *IndicDLP* as a pretraining resource for document layout parsing, observing further performance gains and faster convergence. To facilitate future research on language-specific document understanding, we release *IndicDLP* along with trained models, datasets, and evaluation scripts.

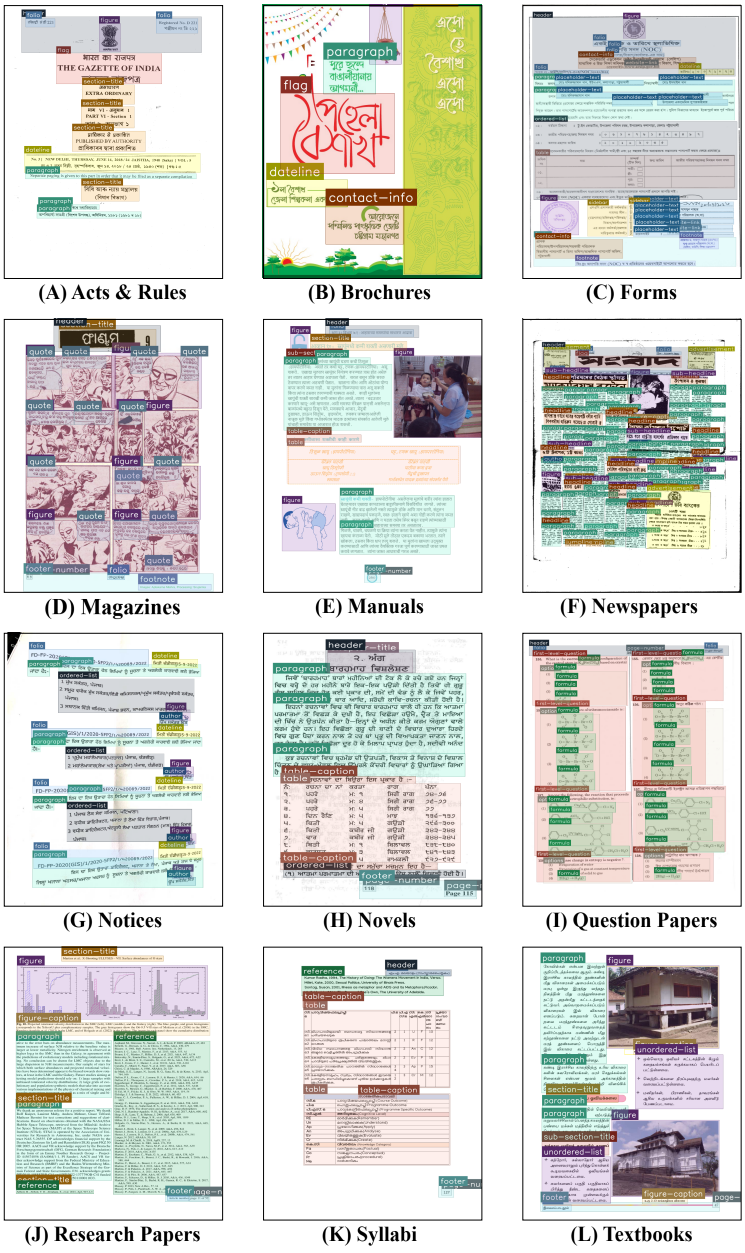

Fig. 1. Samples from the *IndicDLP* dataset highlighting its diversity across document formats, domains, languages, and temporal span. For improved differentiability, segmentation masks are used instead of bounding boxes to highlight regions more effectively. This figure is best viewed when zoomed in.

Dataset	#Images	#Region Classes	Annotation Method	#Domains	#Languages
PRImA [3]	1,240	10	Automatic	5	1
PubLayNet [32]	360,000	5	Automatic	1	1
DocBank [18]	500,000	13	Automatic	1	1
DocLayNet [22]	80,863	11	**Manual**	6	4
M^6*Doc* [10]	9,080	**75**	**Manual**	7	2
D^{4}LA [11]	11,092	27	**Manual**	12	1
BaDLAD [24]	33,695	4	**Manual**	6	1
***IndicDLP*(Ours)**	**121,198**	42	**Manual**	**12**	**12**

Table 1. Comparison of modern document layout parsing datasets.

2 Related Work

Layout Parsing Datasets. The inherent challenges of deep learning-based document parsing have prompted the development of various datasets, each with distinct limitations. Large-scale datasets like PubLayNet [32] and DocBank [18] offer ample size but are constrained by limited region labels and domain diversity. The RVL-CDIP dataset [14] improves domain diversity for document classification tasks but uses grayscale images, restricting its relevance for modern layouts. Similarly, D^{4}LA [11], a manually annotated subset of RVL-CDIP designed for layout parsing, faces the same limitation. Datasets such as *M*^{6}*Doc* [10], DocLayNet [22], and BaDLAD [24] mitigate some of these issues by incorporating diverse domains and document layouts, but still lack source variety and suffer from overly fine-grained labels. Moreover, *M*^{6}*Doc* [10] reports poor cross-dataset performance between *M*^{6}*Doc* and DocLayNet [22], highlighting the need for a more diverse dataset that covers a broad range of domains and provides an adequate representation of common document types. As we show in Section 3, our dataset *IndicDLP* addresses these shortcomings, offering broader coverage and better representation of document types.

Layout Parsing Models. Existing research on layout parsing has primarily focused on visual approaches, advancing document understanding through models like RCNNs [23,15,8] and single-shot detectors such as SSD [21] and the YOLO series [26]. Transformers further improved layout parsing performance with models like DocSegTr [7], SwinDocSegmenter [6], and TransDLANet [10]. While DiT [17] leveraged self-supervised pretraining, DocLayout-YOLO [31] used synthetic pretraining for competitive results. RoDLA [9] introduced robustness enhancements via a variant of DINO [29]. More recently, multimodal models incorporating textual, visual, and spatial features [30,4,25,19] have emerged, but their effectiveness in modality integration is limited by the scarcity of multilingual textual data, especially for Indian languages with limited OCR resources. Therefore, this work focuses on visual document layout parsing models to highlight state-of-the-art performance on multidomain Indic document data.

3 *IndicDLP*

3.1 Motivation

To build useful and general layout detection models, a document layout dataset should (i) be large in terms of document count (ii) span a diverse and representative set of domain categories (e.g. newspapers, forms, textbooks) (iii) provide a comprehensive set of region labels with sufficient label frequency (iv) cover a large set of languages. For the remainder of this section, we describe our dataset and demonstrate how it meets the above criteria.

At first glance, the language coverage criteria do not seem very obvious since layout detection appears to be script agnostic. However, many Indic languages have intricate scripts characterized by inflections, diacritics, and composite or conjugated characters which complicates layout detection (see Figure 1). Also, a key feature of many Indic scripts is that the visual form of certain characters in a word changes based on their interaction with neighboring letters. These script-level complexities are further amplified by the prevalence of scanned copies or photographs as the primary form of publicly available Indic documents, limiting the effectiveness of synthetic methods for automatic annotation. In addition, many printed Indic documents span multiple decades in origin and exhibit significant typographical and printing variations. These historical texts pose challenges due to their complex layouts, frequently illegible writing, deteriorated or stained paper quality, and non-standard formatting, making automated processing especially difficult.

From a downstream application perspective, poor localization negatively affects the performance of transcription pipelines for Optical Character Recognition (OCR). To address this, a high-quality and robust general layout parsing model trained on manually-annotated diverse Indic documents, spanning multiple scripts, domains and formats, is required.

3.2 Dataset domains and sources

IndicDLP comprises 121,198 layout-annotated document images in 11 different Indian languages and English. These documents span 12 document domain categories: newspapers (5.7%), novels (11%), magazines (10.6%), manuals (9.6%), forms (5.6%), textbooks (10.9%), acts and rules (10%), question papers (6.5%), brochures (8.4%), notices (5.4%) and syllabi (6.8%), and research papers (9.4%). The dataset covers a wide range of formats, including documents from the pre-independence era to the present day.

The document images are sourced from various digital-born files, scanned documents, or photographed pages. In total, they include 1,856,241 annotated instances, encompassing 42 physical and logical region labels. In the following sections, we detail our data curation process (Section 3.3), our methodology for selecting region labels (Section 3.4) and the annotation and review workflow (Section 3.5).

Domain	#Inst	%	ARD	AAT(s)
Acts & Rules **(ar)**	12,210	10.074	10	766
Brochures **(br)**	10,128	8.357	14	1207
Forms **(fm)**	6,751	5.570	23	1196
Magazines **(mg)**	12,807	10.567	15	1667
Manuals **(mn)**	11,660	9.621	10	330
Newspapers **(np)**	6,886	5.682	72	3682
Notices **(nt)**	6,602	5.447	11	354
Novels **(nv)**	13,385	11.044	10	304
Question Papers **(qp)**	7,866	6.490	17	708
Research Papers **(rp)**	11,435	9.435	9	322
Syllabi **(sy)**	8,222	6.784	11	445
Textbooks **(tb)**	13,246	10.929	11	520
Total	**121,198**	**100**		

Table 2. Overview of the different domains in *IndicDLP* , including the total number of documents, percentage contribution to *IndicDLP*, average number of regions per document [ARD] (rounded to the nearest integer), and average annotation time per document [AAT] (in seconds). The disproportionately high average annotation time for newspapers is due to the large number of regions they typically contain.

3.3 Domain Categories

To ensure dataset diversity, we curated documents encountered in popular use case scenarios. The sources for each category are detailed below, and representative samples across different languages are shown in Figure 1.

Newspapers: This domain includes both modern and historical newspapers. Modern ones are digitally generated or high-quality scans, while historical scans exhibit older fonts, disjoint lettering, and unique glyphs like conjuncts. These documents often feature multi-page, multi-column, multi-section layouts with diverse labels, making them among the most complex domains in the dataset.
Magazines & Manuals: Magazines and manuals typically contain single or double-page multi-column layouts. We include image-rich comics and advertisement magazines from online sources to enhance layout diversity.
Novels: Indic novels offer insights into traditional typesetting and typography. We curated PDFs from various non-copyrighted ebook stores, including covers, tables of contents, and indices wherever available, to increase layout variety.
Textbooks: Textbooks used in Indian schools with native language teaching mediums are typically written in regional languages. They often feature interleaved text and images, contributing to visual and structural diversity.
Acts & Rules: We include digitally generated or scanned public documents and bulletins released by the Indian Government. These documents feature logos, seals, and intricate headers, enriching the variety of layouts.
Question Papers: Indian schools and national exams often provide bilingual question papers in English or Hindi along with a native language. These documents feature rich multi-script regions within a single source, making them valuable for training models to handle script diversity effectively.

Forms: Forms feature structured fields, tables, and handwritten input areas. Our collection includes both government-issued and organizational forms in both digital and scanned formats.
Brochures: Brochures, pamphlets, and leaflets typically feature diverse graphical layouts with sparse text. *IndicDLP* includes brochures from government agencies, educational institutions, and various commercial sources such as retail, healthcare, tourism, finance, and real estate, capturing diverse design styles.
Notices & Syllabi: Structured with headers, lists, and tables, these documents come from schools, universities, online bulletin boards and government sources, offering a wide range of academic and administrative layouts.
Research Papers: Indian research papers, theses, and multidisciplinary scholarly articles are a rich source of structured academic writing, typically formatted in single or double-column layouts with tables, figures, and citations. Our dataset includes research articles downloaded as PDFs from open-access repositories such as Shodhganga [1] and arXiv, capturing diverse academic styles, equations, and multilingual content.

3.4 Region Label Set Curation

We analyzed existing layout parsing datasets to define a concise yet comprehensive label set for document understanding tasks. To ensure broad coverage, we followed *M⁶Doc* [10] guidelines, incorporating region labels for specific document categories. This was crucial for domains like newspapers and forms, which contain unique elements such as *jumplines* and *placeholders*, elements typically absent from other domains.

To establish an initial set of region labels covering over 90% of common document structures, we also reviewed several other layout parsing datasets. Consistent with *M⁶Doc* [10] findings, we observed that datasets like PubLayNet [32], DocBank [18], and DocLayNet [22] primarily focus on basic, domain-independent labels. To achieve finer granularity, we expanded the label set by an additional 31 categories. For example, we split *list* into *ordered* and *unordered*, and *caption* into *table caption* and *figure caption*. Our dataset also includes multi-level hierarchical labels for elements such as *headlines*, *section titles*, and *lists*.

Further refinement was based on a careful manual analysis of our dataset images, focusing on the frequency and visual prominence of specific regions. We excluded rarely occurring labels like *QR code*, *bar code*, and *poem* (included in *M⁶Doc* [10]) to reduce complexity. Instead, we included frequently encountered regions such as *website links*, *contact information*, and *quotations*, which are common in magazines and newspapers. The complete set of region labels and their instance counts are presented in Table 3.

3.5 Annotation and Review Workflow

Our primary objective in this work is to develop a uniform and consistent set of annotated document pages for layout parsing across multiple languages. To

Region Labels	Train		Val		Test	
	#Inst	%	#Inst	%	#Inst	%
advertisement	19175	1.293	2038	1.086	2110	1.134
answer	4100	0.277	449	0.239	556	0.299
author	17860	1.205	2247	1.197	2171	1.167
chapter-title	6471	0.436	833	0.444	781	0.420
contact-info	12070	0.814	1695	0.903	1538	0.827
dateline	31158	2.102	3853	2.053	3925	2.110
figure	76957	5.191	9735	5.187	10022	5.388
figure-caption	25202	1.700	3201	1.706	3396	1.826
first-level-question	37009	2.496	4742	2.527	4838	2.601
flag	3337	0.225	424	0.226	447	0.240
folio	82112	5.538	10501	5.595	10476	5.632
footer	55787	3.763	7068	3.766	6974	3.749
footnote	12304	0.830	1557	0.830	1539	0.827
formula	15157	1.022	1959	1.044	1883	1.012
header	62207	4.196	7867	4.192	7842	4.216
headline	44758	3.019	5557	2.961	5480	2.946
index	460	0.031	81	0.043	33	0.018
jumpline	3897	0.263	544	0.290	512	0.275
options	35575	2.400	4518	2.407	4699	2.526
ordered-list	40672	2.743	5232	2.788	5017	2.697
page-number	79316	5.350	10025	5.342	9933	5.340
paragraph	479891	32.369	60964	32.485	59860	32.183
placeholder-text	77291	5.213	9367	4.991	9708	5.219
quote	3495	0.236	455	0.242	337	0.181
reference	5442	0.367	700	0.373	658	0.354
second-level-question	13203	0.891	1725	0.919	1572	0.845
section-title	81270	5.482	10542	5.617	10123	5.443
sidebar	39868	2.689	5182	2.761	4847	2.606
sub-headline	15226	1.027	1928	1.027	2105	1.132
sub-ordered-list	18795	1.268	2399	1.278	2313	1.244
sub-section-title	17908	1.208	2164	1.153	2333	1.254
sub-unordered-list	866	0.058	90	0.048	70	0.038
subsub-headline	1892	0.128	201	0.107	235	0.126
subsub-ordered-list	4455	0.300	610	0.325	555	0.298
subsub-section-title	2506	0.169	237	0.126	274	0.147
subsub-unordered-list	96	0.006	3	0.002	8	0.004
table	19389	1.308	2407	1.283	2402	1.291
table-caption	7376	0.498	920	0.490	930	0.500
table-of-contents	956	0.064	135	0.072	109	0.059
third-level-question	2669	0.180	265	0.141	252	0.135
unordered-list	12795	0.863	1652	0.880	1676	0.901
website-link	11599	0.782	1598	0.851	1460	0.785
Total	**1482572**	100	**187670**	100	**185999**	100

Table 3. Distribution of layout regions in *IndicDLP*, sorted in alphabetical order of region label names.

achieve this, we trained a team of 50 individuals, comprising 3 to 4 annotators and 1 reviewer per language. We developed a 150-page guideline to ensure clear and consistent region-based annotation, supplemented with annotated examples spanning multiple languages and domains. Rectangular bounding boxes were chosen for their consistency across languages and domains, enabling efficient large-scale annotation. This choice offered a practical trade-off between

the higher precision of polygonal or rotated boxes in skewed cases and overall annotation efficiency. We point out several key annotation decisions from the guidelines that are different from datasets like M^6Doc [10], PubLayNet [32] and DocLayNet [22]:

- Figures include plots, graphs, barcodes, seals, QR codes, and logos.
- Text within figures and sub-figures is *not* annotated separately.
- Variable names in research papers are *not* labeled as formulae to reduce ambiguity, unlike M^6Doc [10]. Clear guidelines specify when mathematical notation should be annotated as formulae.
- Items in the header or footer sections are labeled appropriately and enclosed within a *header* or *footer* region container, respectively.

All annotations were performed using Shoonya [2], an open-source annotation framework based on Label Studio. We meticulously tracked metadata, including annotation performance, average annotation time, draft mode usage, and correction records. Each annotation was validated by a language-proficient reviewer, and a team of 10 supercheckers, including the authors, further checked over 60% of the validated documents to ensure cross-language and cross-domain consistency. The entire annotation process, including two-step validation, was completed in approximately eight months. Finally, for dataset preparation, images were scaled to a maximum of 1024 pixels on the shortest side and 1333 pixels on the longest. The dataset was split into training, validation, and testing sets in an 80:10:10 ratio using a stratified approach to preserve language and domain proportions. The distribution of region labels for each split is presented in Table 3. Further details on language-wise annotated data distribution and annotation workflow can be found in our project page.

4 Experiments

In this section, we present a comprehensive qualitative and quantitative evaluation of various state-of-the-art (SOTA) document layout parsing and object detection models trained on the *IndicDLP* dataset. Object detection is chosen as the primary task for layout parsing, with mean Average Precision (mAP@[.5:.95]) as the performance metric, calculated over thresholds ranging from 0.5 to 0.95 in steps of 0.05. We trained several state-of-the-art models, including YOLOv10 [27], DiT [17], DocLayout-YOLO [31], DINO [29], RoDLA [9], and Florence-2 [28], exploring different model sizes and architectures to establish baselines. Original hyperparameters, as specified in the respective papers, were retained for our experiments. All models were trained and evaluated exclusively on our *IndicDLP* dataset on 8 NVIDIA H100 GPUs. Detailed quantitative results are provided in Table 4.

4.1 How do different models perform on *IndicDLP* ?

As shown in Table 4, the diverse region labels and their varying appearances across domains in *IndicDLP* pose significant challenges for all models. No-

Model	Model Size(M)	mAP50(↑)	mAP75(↑)	mAP[50-95](↑)
YOLOv10x [27]	**37**	73.5	**60.6**	**55.0**
DocLayout-YOLO [31]	20	73.5	60.0	54.5
RoDLA [9]	342	**74.1**	57.7	53.1
DINO [29]	46	69.7	53.4	49.2
DIT [17]	86	67.2	51.8	47.8
Florence-2 [28]	826	41.4	28.7	28.0

Table 4. Performance comparison on *IndicDLP*.

tably, DiT [17], RoDLA [9], and DocLayout-YOLO [31] were the only models pretrained on document-based datasets: IIT-CDIP, *M⁶Doc* [10], and DocSynth-300K [31], respectively. Except for Florence-2 [28], which was pretrained on the FLD-5B dataset, all other models were pretrained on the COCO natural images dataset [20]. Among them, YOLOv10x [27] emerges as one of the top performers for layout parsing. Therefore, we conduct further ablation and fine-tuning analysis on YOLOv10x [27] due to its high efficiency and low memory footprint.

We observe consistent patterns across all evaluated models. For instance, regions requiring semantic understanding, such as *jumpline* (which indicates the continuation of text on another page or section) or *author*, tend to underperform when models rely solely on visual features. Similarly, regions like *index* (list of references at the back of a book) or *subsub-unordered-list* (the third hierarchical level within an unordered list) also perform poorly due to their relatively low frequency in the dataset. In contrast, regions like *placeholder-text* (text entry fields) and *page-number*, though more frequent, show lower performance due to their diverse visual appearances. On the other hand, regions like *footnote* and *flag* (the nameplate/branding of a newspaper/magazine) demonstrate higher accuracy despite their lower frequency, likely due to their consistent appearance and fixed positions, usually at the bottom or top of the document page, respectively.

On *IndicDLP*, YOLOv10x [27] and DocLayout-YOLO [31] perform similarly, likely because DocLayout-YOLO [31] is a variant of the former with document-specific optimizations in its design. RoDLA [9], intended to be a more robust version of DINO [29], outperforms DINO, indicating its effectiveness on naturally degraded document images in our dataset. Despite its significantly larger capacity, Florence-2 [28] underperforms compared to specialized object detection models, likely due to its pretraining emphasis on diverse natural images rather than document-specific layouts. For further qualitative insights, we present sample predictions from the YOLOv10x model, trained from scratch on DocLayNet, D⁴LA, *M⁶Doc*, and *IndicDLP*. These predictions are compared against the Ground Truth annotations in Figure 3.

4.2 How do models perform across the different domains in *IndicDLP*?

The diversity of document styles across domains in *IndicDLP*, as illustrated in Figure 1, leads to substantial variation in domain-wise performance. We show the

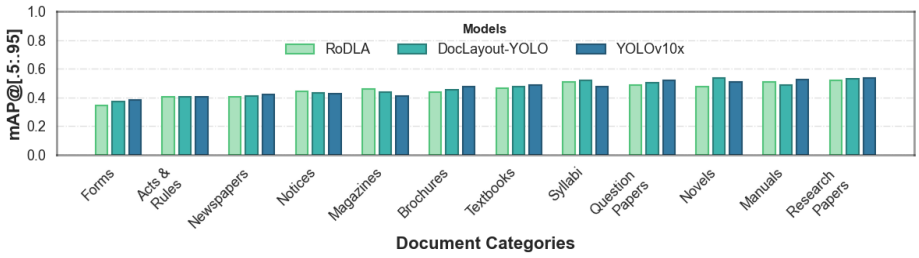

Fig. 2. Comparison of mAP Scores Across Different Domains for YOLOv10x, DocLayout-YOLO and RoDLA trained on *IndicDLP*.

domain-wise performance of the top 3 best performing models on *IndicDLP* in Figure 2. We note that domains such as *Forms* and *Acts & Rules* have significantly different layouts compared to others, negatively impacting the performance of the model. Certain domains like *Newspapers* and *Brochures* contain a wider range of unique regions to identify, making precise region recognition more complex. Additionally, the dense, multi-column layouts found in *Newspapers* and *Magazines* also make detection tasks more challenging. In contrast, categories with simpler, more uniform, and predictable layouts, such as *Novels* and *Research Papers*, achieve higher mAP scores due to better detection accuracy.

4.3 How do models perform across the different languages in *IndicDLP*?

Indic scripts across the nation vary significantly in structure and character complexity. The multilingual nature of *IndicDLP* offers a valuable opportunity to study how script differences affect layout parsing performance. To support language-specific evaluation, we split *IndicDLP* into subsets by language and fine-tune a YOLOv10x model for each. This experimental setup also allows us to assess cross-lingual transfer in layout parsing, particularly between languages with shared scripts or from geographically proximate regions. Specifically, we train on documents in one script and evaluate performance on documents in another.

Quantitative results of this experiment are presented in Table 5. Comparing the diagonal 'same script' entries with those in the last row (i.e. full *IndicDLP*), we find that a model trained on all scripts outperforms those trained specifically for each script. The substantially larger training set of the full *IndicDLP* trained model compensates for the lack of script homogeneity. To isolate the impact of dataset size, we create a smaller *IndicDLP*-mini subset, consisting of 10K images proportionally sampled from all languages of *IndicDLP* while maintaining consistent domain distribution within each language. The relatively higher performance of the diagonal entries compared to the *IndicDLP*-mini baseline (second-to-last row in Table 5) suggests that script influences layout prediction when training data is adequately accounted for.

↓ Train \ Test →	en	hi	mr	gu	pa	bn	as	or	ta	te	kn	ml
English	0.39	0.18	0.18	0.12	0.18	0.17	0.13	0.15	0.18	0.15	0.16	0.19
Hindi	0.18	0.45	0.27	0.19	0.20	0.23	0.23	0.19	0.18	0.19	0.21	0.20
Marathi	0.16	0.28	0.55	0.19	0.17	0.23	0.24	0.20	0.19	0.14	0.20	0.18
Gujarati	0.16	0.20	0.23	0.42	0.17	0.20	0.22	0.19	0.18	0.14	0.18	0.16
Punjabi	0.13	0.20	0.21	0.18	0.49	0.20	0.21	0.18	0.15	0.13	0.16	0.16
Bengali	0.14	0.21	0.23	0.16	0.17	0.46	0.26	0.22	0.18	0.16	0.18	0.17
Assamese	0.13	0.23	0.24	0.14	0.20	0.27	0.55	0.23	0.18	0.14	0.17	0.19
Odiya	0.12	0.18	0.20	0.14	0.15	0.20	0.20	0.49	0.16	0.15	0.17	0.17
Tamil	0.14	0.17	0.18	0.12	0.15	0.18	0.17	0.18	0.49	0.18	0.21	0.20
Telugu	0.11	0.14	0.14	0.11	0.14	0.16	0.14	0.14	0.16	0.43	0.18	0.18
Kannada	0.18	0.22	0.21	0.13	0.16	0.18	0.16	0.18	0.22	0.20	0.44	0.20
Malayalam	0.18	0.15	0.19	0.12	0.15	0.17	0.16	0.19	0.21	0.16	0.21	0.47
IndicDLP-mini	0.32	0.37	0.45	0.32	0.36	0.38	0.46	0.40	0.41	0.32	0.34	0.42
IndicDLP	**0.50**	**0.52**	**0.60**	**0.47**	**0.58**	**0.53**	**0.63**	**0.59**	**0.60**	**0.49**	**0.50**	**0.55**

Table 5. Zero-shot performance evaluation of YOLOv10x trained on language subsets of *IndicDLP*. Language codes follow the **ISO 639-2** standard. Highlighted languages share the same script: *Devanagari* for (**hi**, **mr**) and *Bengali* for (**as**, **bn**). The model trained on the full *IndicDLP* dataset (last row) significantly outperforms those trained monolingually, benefiting both from a larger data volume and cross-lingual transfer.

Pretraining Dataset(PD)	PD Size	mAP50	mAP75	mAP[50:95]
UED-mini	75K	76.4	63.7	**57.7** (+1.9)
No Pretraining (Scratch)	–	74.5	61.3	55.8 (Baseline)
COCO	118K	73.5	60.6	55.0 (**-0.8**)
UED	785K	72.8	59.4	53.9 (**-1.9**)
PubLayNet	360K	71.7	59.2	53.0 (**-2.8**)

Table 6. Performance of YOLOv10x pretrained on different sources and fine-tuned on *IndicDLP*. Pretraining on smaller, diverse datasets like UED-mini yields better results than larger generic datasets.

From Table 5, the cross-diagonal results show that zero-shot evaluation on unseen languages generally leads to mAP scores that are 20–25 points lower. This suggests that models trained on a single language rely heavily on script-specific features and generalize poorly to unseen scripts. This, along with the results in the last row, highlights the benefits of incorporating multiple languages in foundational layout datasets. Interestingly, while cross-lingual transfer improves slightly for languages sharing the same script (marked in **red** for *Devanagari* and **blue** for *Bengali* in Table 5), non-script-based visual differences (i.e., distribution shift) lead to cross-lingual models for shared-script languages underperforming compared to language-specific models.

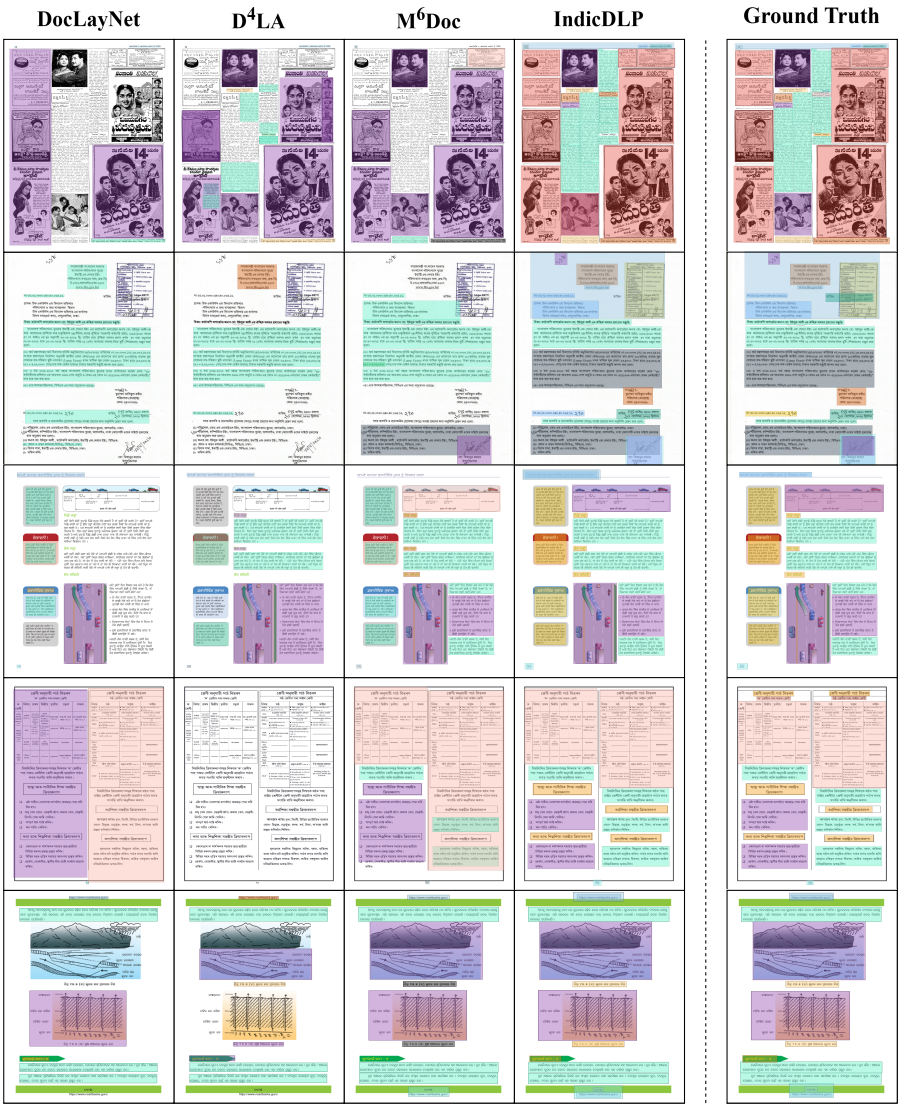

Fig. 3. Qualitative analysis of YOLOv10x model predictions on *IndicDLP* test set when trained from scratch on DocLayNet (Column 1), D^{4}LA (Column 2), *M^{6}Doc* (Column 3), and *IndicDLP* (Column 4), compared to the Ground Truth (Column 5). For instance, in Rows 1 and 3, for *M^{6}Doc*, despite having labels like *advertisements* and *sidebars*, distribution differences prevent accurate localization and classification. In Rows 2 and 4, models trained on *IndicDLP* show superior performance on non-standard and multicolumn layouts, respectively. Row 5 highlights partial *figure* detections for DocLayNet and D^{4}LA, even though it is a common region across all the above datasets.

4.4 Does English pretraining improve layout parsing performance?

Our preliminary experiments indicate that training on high-quality datasets specific to layout parsing can further improve fine-tuning performance on *IndicDLP*. This strategy is particularly effective due to the considerable variability in quantity, domain, document types, and visual appearances across existing English-dominant layout parsing datasets. In the following section, we explore which data configuration is more effective: (i) using all available datasets, which provides a larger volume of data, or (ii) curating high-quality, manually annotated subsets, which offer greater data diversity.

To test our approach, we evaluated several layout parsing datasets: PubLayNet [32], DocBank [18], DocLayNet [22], *M⁶Doc* [10], and D⁴LA [11], creating two bundles: UED (Unified Existing Datasets) and a curated subset, UED-mini (see Table 6). UED-mini prioritized human-annotated datasets with diverse domains and logical layout elements. PubLayNet and DocBank were excluded due to automated annotations, and D⁴LA for its grayscale images with limited inter-domain variability. DocLayNet and *M⁶Doc* had inconsistently annotated or underrepresented regions, limiting generalization. We refined and merged their labels iteratively, consolidating semantically similar categories (e.g., *table-caption* and *image-caption* as *caption*; *text* in DocLayNet mapped to *paragraph* in *M⁶Doc*). Less significant labels were removed, resulting in a unified taxonomy of 25 labels. This formed the UED-mini dataset, combining the strengths of both the source datasets.

To assess the impact of pretraining, we compared YOLOv10x trained from scratch on *IndicDLP* with variants pretrained on UED, UED-mini, COCO [20], and PubLayNet, followed by fine-tuning on *IndicDLP*. Results are presented in Table 6.

We observe that pretraining on English datasets does not consistently improve performance. UED, dominated by synthetic data from PubLayNet and DocBank (700K of 785K images), suffers from limited diversity, which impacts performance more than the domain gap introduced by COCO. In contrast, models pretrained on the curated UED-mini set perform better, indicating the necessity of careful selection and curation of existing datasets. These results highlight the value of both domain diversity and high-quality annotations, with UED-mini proving to be the most effective pretraining source for *IndicDLP*.

4.5 Conversely, how effective is *IndicDLP* as the initial pretraining dataset?

In this section, we evaluate the effectiveness of *IndicDLP* as a pretraining dataset for layout parsing in other languages. For this, we pretrain two models: one using *IndicDLP* and the other using COCO (as the baseline). Both models are then fine-tuned on *M⁶Doc* [10], DocLayNet [22], and D⁴LA [11].

As shown in Figure 4, using *IndicDLP* for pretraining offers notable advantages: it either accelerates convergence to optimal values (as observed with DocLayNet [22]), yields significant performance improvements (as seen with

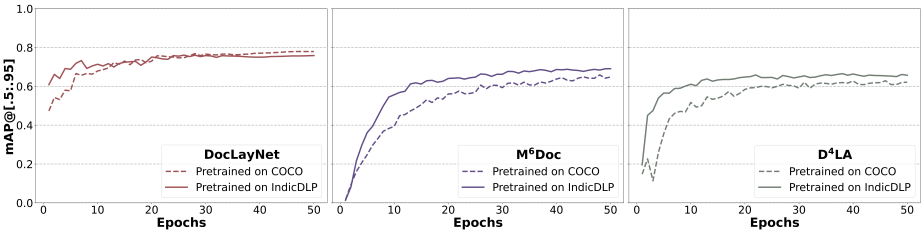

Fig. 4. Performance gap observed during finetuning on various datasets using YOLOv10x pretrained on *IndicDLP* compared to finetuning from scratch. The solid line represents finetuning with *IndicDLP*, while the dashed line represents finetuning without it. For *M⁶Doc* (center) and D⁴LA (right), pretraining on *IndicDLP* increases mAP by 2.8 points on average, while for DocLayNet (left), it leads to faster convergence.

M⁶Doc [10]), or achieves both (in the case of D⁴LA [11]), demonstrating the effectiveness of *IndicDLP* as a good pretraining dataset for document layout parsing tasks in other languages.

5 Conclusion

Document layout parsing is critical for advancing document understanding, yet existing datasets suffer from limitations in scale, diversity, and annotation granularity. *IndicDLP* fills this gap by providing the largest and most diverse dataset for Indic document layout analysis, spanning 12 languages, 12 domains, and 42 region labels. We also introduce UED-mini as a curated dataset and demonstrate its utility for pretraining. Our experiments demonstrate that models trained on *IndicDLP* generalize well across a wide range of document layouts, improving both Indic and non-Indic document parsing. Our work opens up avenues for enhancing OCR, retrieval, and information extraction for Indic scripts, thereby facilitating inclusive and high-performing document digitization solutions. We release *IndicDLP* as an open resource dataset, enabling the community to push the boundaries of multilingual, multi-domain document understanding.

6 Acknowledgements

This work is supported by the Ministry of Electronics and Information Technology (MeitY), Government of India, as part of the Digital India Bhashini Mission, which aims to advance Indian language technology. The project's human resources, including annotators, reviewers, and supercheckers, were supported through a dedicated grant provided to IIT Madras, which serves as the Data Management Unit for the mission.

References

1. Shodhganga, available at: https://shodhganga.inflibnet.ac.in/ 3.3
2. Shoonya, available at: https://github.com/AI4Bharat/Shoonya 3.5
3. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: Prima: A realistic dataset for performance evaluation of document layout analysis. In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009). pp. 296–300. Barcelona, Spain (2009) 2
4. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2
5. Auer, C., Dolfi, M., Carvalho, A., Berrospi Ramis, C., Staar, P.W.J.: Delivering Document Conversion as a Cloud Service with High Throughput and Responsiveness. In: Proceedings of the 2022 IEEE 15th International Conference on Cloud Computing (CLOUD). pp. 363–373. IEEE (2022). https://doi.org/10.1109/CLOUD55607.2022.00060 1
6. Banerjee, A., Biswas, S., Lladós, J., Pal, U.: Swindocsegmenter: An end-to-end unified domain adaptive transformer for document instance segmentation. In: Document Analysis And Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part I. Lecture Notes in Computer Science, vol. 14187, pp. 307–325 (2023). https://doi.org/10.1007/978-3-031-41676-7_18 2
7. Biswas, S., Banerjee, A., Lladós, J., Pal, U.: Docsegtr: An instance-level end-to-end document image segmentation transformer (2022) 2
8. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection (2017) 2
9. Chen, Y., Zhang, J., Peng, K., Zheng, J., Liu, R., Torr, P., Stiefelhagen, R.: Rodla: Benchmarking the robustness of document layout analysis models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 1, 2, 4, 4.1, 4.1
10. Cheng, H., Zhang, P., Wu, S., Zhang, J., Zhu, Q., Xie, Z., Li, J., Ding, K., Jin, L.: M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15138–15147 (2023) 1, 2, 2, 3.4, 3.5, 4.1, 4.4, 4.5
11. Da, C., Luo, C., Zheng, Q., Yao, C.: Vision grid transformer for document layout analysis. In: IEEE International Conference on Computer Vision (ICCV) (2023) 1, 2, 4.4, 4.5
12. Fateh, A., Fateh, M., Abolghasemi, V.: Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection. Engineering Reports **6**(9), e12832 (2024). https://doi.org/10.1002/eng2.12832 1
13. Ha, H.T., Horák, A.: Information Extraction from Scanned Invoice Images using Text Analysis and Layout Features. Signal Processing: Image Communication **102**, 116601 (2022). https://doi.org/10.1016/j.image.2021.116601 1
14. Harley, A., Ufkes, A., Derpanis, K.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR) (2015) 2
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn (2017) 2
16. Jaha, E.S.: Comparative Semantic Document Layout Analysis for Enhanced Document Image Retrieval. IEEE Access **12**, 150451–150467 (2024). https://doi.org/10.1109/ACCESS.2024.3479990 1

17. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: Dit: Self-supervised pre-training for document image transformer. In: Proceedings of the ACM Multimedia Conference (2022) 1, 2, 4, 4.1
18. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: Docbank: A benchmark dataset for document layout analysis. In: Proceedings of the International Conference on Computational Linguistics (COLING) (2020) 1, 2, 3.4, 4.4
19. Li, P., Gu, J., Kuen, J., Morariu, V.I., Zhao, H., Jain, R., Manjunatha, V., Liu, H.: Selfdoc: Self-supervised document representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2
20. Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV) (2014) 4.1, 4.4
21. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.: Ssd: Single shot multibox detector (2015) 2
22. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A., Staar, P.: Doclaynet: A large human-annotated dataset for document-layout segmentation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3743–3751 (2022) 1, 2, 3.4, 3.5, 4.4, 4.5
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks (2015) 2
24. Shihab, M., Hasan, M., Emon, M., Hossen, S., Ansary, M., Ahmed, I., Rakib, F., Dhruvo, S., Dip, S., Pavel, A., Meghla, M., Haque, M., Chowdhury, S., Sadeque, F., Reasat, T., Humayun, A., Sushmit, A.: Badlad: A large multi-domain bengali document layout analysis dataset. In: IEEE International Conference on Document Analysis and Recognition (ICDAR) (2023) 2
25. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., Bansal, M.: Unifying vision, text, and layout for universal document processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2
26. Terven, J., Cordova-Esparza, D.: A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. ArXiv (2023) 2
27. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: Yolov10: Real-time end-to-end object detection. In: Proceedings of the Neural Information Processing Systems Conference (NeurIPS) (2024) 1, 4, 4.1, 4.1
28. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 1, 4, 4.1, 4.1
29. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. ArXiv (2022) 1, 2, 4, 4.1, 4.1
30. Zhang, N., Cheng, H., Chen, J., Jiang, Z., Huang, J., Xue, Y., Jin, L.: M2doc: A multi-modal fusion approach for document layout analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7233–7241 (2024) 2
31. Zhao, Z., Kang, H., Wang, B., He, C.: Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. ArXiv (2024) 1, 2, 4, 4.1, 4.1
32. Zhong, X., Tang, J., Jimeno-Yepes, A.: Publaynet: Largest dataset ever for document layout analysis. In: IEEE International Conference on Document Analysis and Recognition (ICDAR) (2019) 1, 2, 3.4, 3.5, 4.4