OMNILAYOUT: ENABLING COARSE-TO-FINE LEARNING WITH LLMS FOR UNIVERSAL DOCUMENT LAYOUT GENERATION

Anonymous authors
Paper under double-blind review

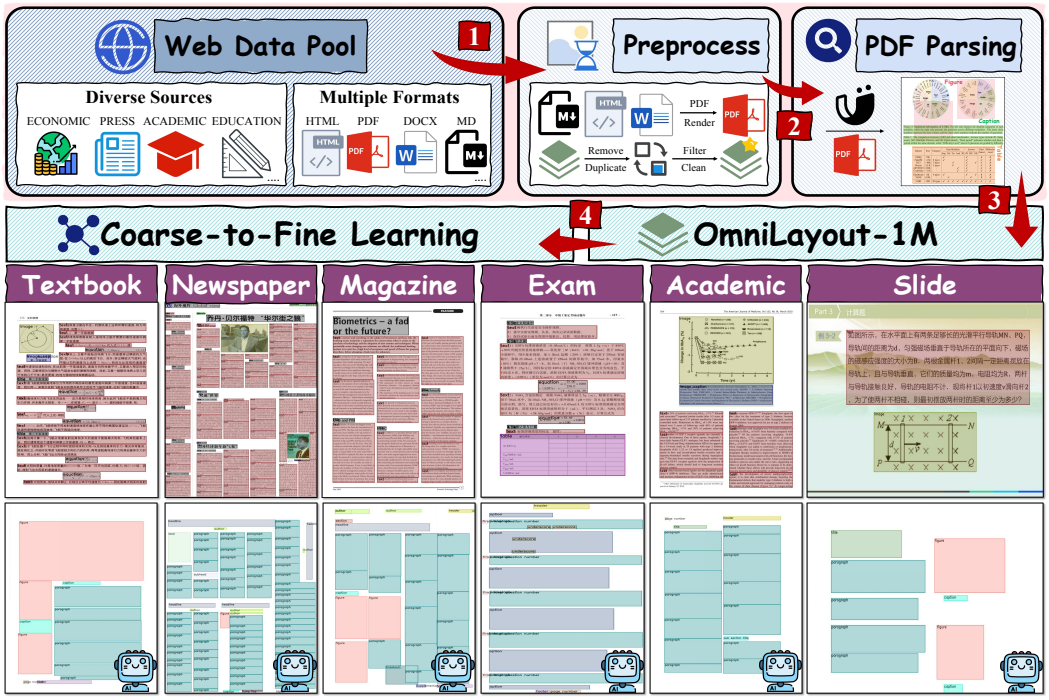


Figure 1: **Overview of OmniLayout.** (Top & Middle) show the curation and examples of OmniLayout-1M. (Bottom) illustrates unconditional layouts generated by our OmniLayout-LLM via coarse-to-fine learning.

ABSTRACT

Document AI has advanced rapidly and is attracting increasing attention. Yet, while most efforts have focused on document layout analysis (DLA), its generative counterpart, document layout generation, remains underexplored. A major obstacle lies in the scarcity of diverse layouts: academic papers with Manhattan-style structures dominate existing studies, while open-world genres such as newspapers and magazines remain severely underrepresented. To address this gap, we curate OmniLayout-1M, the first million-scale dataset of diverse document layouts, covering six common document types and comprising contemporary layouts collected from multiple sources. Moreover, since existing methods struggle in complex domains and often fail to arrange long sequences coherently, we introduce OmniLayout-LLM, a 0.5B model with designed two-stage *Coarse-to-Fine learning paradigm*: 1) learning universal layout principles from OmniLayout-1M with coarse category definitions, and 2) transferring the knowledge to a specific domain with fine-grained annotations. Extensive experiments demonstrate that our approach achieves strong performance on multiple domains in M^{6}Doc dataset, substantially surpassing both existing layout generation experts and several latest general-purpose LLMs. Our code, models, and dataset will be publicly released.

1 INTRODUCTION

Document AI has attracted growing attention across academia and industry recently, as it plays a critical role in enabling machines to understand, process, and generate documents. On the one hand, increasing efforts have been devoted to document parsing (Wang et al., 2024; Li et al., 2025; Cui et al., 2025), which aims to extract structural and semantic information from massive amounts of pages through layout analysis and optical character recognition (OCR), On the other hand, however, its generative counterpart—document layout generation (Gupta et al., 2021; Kong et al., 2022), has not yet been fully explored. This task focuses on producing well-organized layouts by arranging visual elements like text blocks, tables, and figures in a coherent manner, with great potential for applications including content-driven layout design and document image generation.

A few studies have started investigating document layout generation with different generative models, ranging from early GAN-based approaches (Kikuchi et al., 2021) to more recent diffusion or flow-based methods (Guerreiro et al., 2024). Subsequently, the rise of large language models (LLMs) has opened new possibilities for conditional layout generation, drawing on their extensive prior knowledge and long-context understanding abilities. However, a thorough review of previous studies and datasets reveals notable limitations:

(*i*) *Data Scarcity of Diverse Document Layouts.* The domain bias in existing public datasets poses a critical obstacle to the development of general Document AI. Widely-used datasets such as PubLayNet (Zhong et al., 2019) and DocBank (Li et al., 2020) offer massive annotations but primarily focus on a single domain—always academic articles with relatively simple Manhattan layouts. Although datasets like DocLayNet (Pfitzmann et al., 2022) and D^4LA (Da et al., 2023) include a variety of document types, many of these (*e.g.*, letter) are no longer commonly seen in modern real-world scenarios, and their data sources are often outdated. Among existing resources, M^6Doc (Cheng et al., 2023) and OmniDocBench (Ouyang et al., 2025) stand out as the most valuable datasets to date, as they cover a broader spectrum of contemporary document types, even including highly complex layouts such as newspapers. Unfortunately, they contain only a limited number of samples, making them insufficient to support large-scale training. Overall, the landscape of publicly accessible document data exhibits a severe long-tail distribution: academic articles are overrepresented, while complicated, non-Manhattan layouts such as textbooks remain drastically underrepresented.

(*ii*) *Challenges in Complex and Long-Sequence Scenarios.* Due to the lack of diverse layout data, most existing methods are restricted to simple, homogeneous academic layouts, where progress has plateaued. In contrast, most real-world document layouts are more complex, with finer-grained element categories and a larger number of bounding boxes. Our experiments (Table xx) show that these methods struggle in such settings, especially with long-sequence modeling. Diffusion-based layout generation models like LayoutDM (Inoue et al., 2023) and LACE (Chen et al., 2024) are particularly data-hungry and require extensive training to converge in complex domains. While recent LLM-based conditional layout generation approaches, such as RAG (Wu et al., 2025), CoT (Shi et al., 2025), and in-context learning (Lin et al., 2023), offer promise, direct fine-tuning on complex domains increases learning difficulty and leads to frequent failures. Domain-agnostic models like LayoutNUWA (Tang et al., 2023) and LGGPT (Zhang et al., 2025) represent early progress, but are only tested on limited document types and require substantial computational resources.

To this end, we introduce OmniLayout-1M, the first million-scale dataset for diverse document layout generation. ❶ It contains twice as many samples as DocBank, ❷ covers six common document types from real-world scenarios, and ❸ adopts a fully automated annotation pipeline, providing a powerful foundation for training layout generation models. Moreover, to enable diverse document layout generation under limited fine-grained annotated data, we propose a unified framework that formulates the task as a two-stage *Coarse-to-Fine learning paradigm*. Specifically, we first let an LLM learn basic layout principles such as alignment and spatial organization on OmniLayout-1M across sufficiently diverse document types with coarse-grained labels. Then, with only a small amount of fine-grained annotated data, we perform fine-grained adaptation on a specific domain, enabling controllable and adaptable layout generation with minimal supervision and parameter footprint. Our contribution is summarized as follows:

- We introduce **OmniLayout-1M**, the first million-scale document layout dataset, comprising six commonly used document types and annotations for ten block-level element categories.

- We propose **OmniLayout-LLM**, trained with a *Coarse-to-Fine learning paradigm*, where aesthetic rules are first acquired from diverse layouts and subsequently adapted to specific document domain with fine-grained labels. To the best of our knowledge, we are the first to extend document layout generation to complex and challenging domains such as newspapers.
- Extensive experiments across multiple domains demonstrate that our method achieves state-of-the-art (SOTA) performance consistently. In addition, our visualization examples demonstrate alignment with both aesthetic principles and user expectations.

2 OMNILAYOUT-1M DATASET

2.1 MOTIVATION

Despite the rapid advances in document parsing that have given rise to a variety of document layout datasets in recent years, we observe that existing resources still suffer from several notable limitations. ***(i)* Limited Diversity.** Early document layout datasets, such as PubLayNet (Zhong et al., 2019) and DocBank (Li et al., 2020), are largely derived from large-scale academic paper repositories (*e.g.*, PubMed, arXiv) and thus consist of single-domain pages with relatively simple Manhattan layouts. ***(ii)* Deficient Volume.** Generative tasks typically require more data than detection tasks, particularly for diffusion-based models. However, existing diverse datasets like M^{6}Doc (Cheng et al., 2023) and OmniDocBench (Ouyang et al., 2025), contain samples on the order of $10^2 \sim 10^3$ per document type, making them inadequate for training layout generation models. ***(iii)* Outdated Source.** As document layouts evolve toward improved aesthetics , the timeliness of data sources is critical. Although D^{4}LA (Da et al., 2023) covers 12 document types, its images are sourced from RVL-CDIP (Harley et al., 2015), which contains obsolete formats (*e.g.*, handwritten letters) and consists largely of noisy or skewed scans, substantially degrading the quality of the layout. ***(iv)* Inefficient Annotation.** Recent datasets like DocLayNet (Pfitzmann et al., 2022), often rely on labor-intensive manual annotation, which hinders scalability. With the rapid advancement of document parsing tools (*e.g.*, MinerU (Wang et al., 2024)), returning to fully automated pipelines for accurate layout annotation has become feasible and convenient.

2.2 DATASET CONSTRUCTION

To address the limitations outlined in Section 2.1, we present OmniLayout-1M, the first million-scale document layout dataset, featuring diverse common document types, up-to-date data collected from multiple databases and websites, and a fully automated annotation and filtering pipeline.

Preprocessing. To ensure the diversity of OmniLayout-1M, we collect documents from massive sources on the Internet. During the preprocessing stage, we use format standardization techniques to handle different document formats including PDF, DocX, Markdown, etc. Meanwhile, methods such as deduplication and document quality analysis are employed to filter out noisy documents and ensure the high quality of OmniLayout-1M. Finally, we collect data from 36 sources in total, including Academic Databases (13 sources), Publishers (7 sources), and Document-sharing Platforms (16 sources), covering various fields, such as academia, education, news, economics and etc.

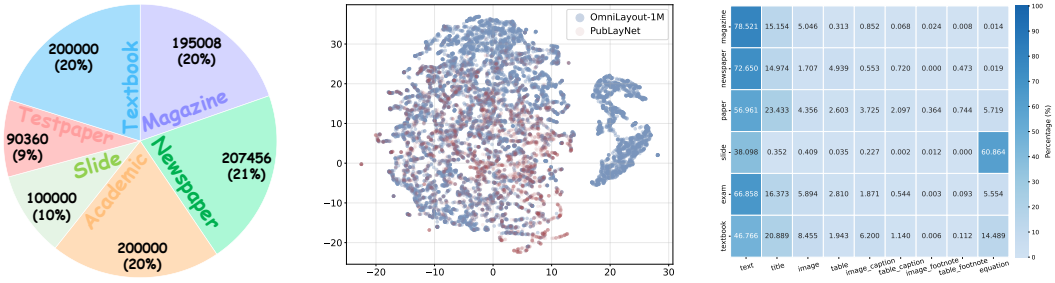
Annotation. To accurately convert the document image into a corresponding element sequence, we employ MinerU (Wang et al., 2024), a powerful open-source toolkit, to automatically annotate the samples. Furthermore, MinerU outputs the element sequence more aligned with the natural reading order, a property essential for reliable and coherent layout generation. In particular, for newspapers whose layouts are quite complicated and that MinerU cannot handle well, we manually annotate 1,000 in-domain newspaper pages and fine-tune a DocLayout-YOLO (Zhao et al., 2024) to produce better performance, especially in capturing dense and irregular layouts.

Dataset	Volume	Element Number	Layout Type	Annotation Method	Source Count
DSSE-200	200	2,546	2	Automatic	Unknown
Prima-LAD	478	7,453	5	Automatic	Unknown
PubLayNet	∼360K	∼3.3M	1	Automatic	1
DocBank	∼500K	∼6.7M	1	Automatic	1
DocLayNet	∼80.9K	∼1.1M	6	Manual	Unknown
D^{4}LA	∼11.1K	∼294K	12	Manual	1
M^{6}Doc	∼9.1K	∼237K	7	Manual	≥ 3
OmniLayout-1M (Ours)	**∼1M**	**∼48.0M**	**6**	**Automatic**	**36**

Table 1: **Comparison with Existing Layout Datasets.**

2.3 Dataset Statistical Analysis

Comparisons with existing datasets. Table 1 highlights OmniLayout-1M's advantages over existing datasets. In terms of diversity, OmniLayout-1M significantly surpasses existing datasets in the number of document elements (about 48M), layout types (6 types), data volume (1M), and data sources diversity. The comprehensive diversity of OmniLayout-1M effectively meets the demand for synthesizing realistic and various document layouts.



(a) Document types distribution. (b) Hand-crafted feature distribution (c) Co-occurrence heatmap

Figure 2: **Statistical Analysis of OmniLayout-1M.**

Document type distribution. The distribution of documents across different layout types is shown in Fig. 2a. OmniLayout-1M encompasses six layout types: textbook, newspaper, magazine, exam, academic, and slide. Data is balanced across all layout types to ensure robust performance.

Hand-crafted feature distribution. OmniLayout-1M exhibits a significantly more layout diversity than simple and homogeneous distribution of academic papers in PubLayNet, as evidenced by an UMAP visualization as shown in Fig. 2b. Hand-crafted features such as number of elements, average area, element centroids are used for visualization.

Element co-occurrence analysis. To validate the plausibility of OmniLayout-1M, Fig. 2c visualizes element co-occurrence patterns. The distributions align with expectations: text and title are most frequent across all document types, followed by image and table, while formula is prominent in academic content such as textbook, paper, and exucational slide. These observations confirm the OmniLayout-1M's adherence to real-world principles.

3 Methodology

3.1 Problem Formulation

Following previous work (Inoue et al., 2023; Guerreiro et al., 2024; Zhang et al., 2025), a document layout is represented as a set of 5-tuples:

$$\mathcal{L} = \{\, e_i = (c, x, y, w, h) \mid i = 1, \ldots, N \,\}, \tag{1}$$

where c denotes the element category, x, y are the horizontal and vertical coordinates of the bounding box (either the top-left corner or the center point), and w, h are its width and height.

As stated in our contributions, unlike all prior approaches, we formulate complex layout generation task as a two-stage *Coarse-to-Fine learning paradigm*, which enables enable the model to learn complex layout logic from easy to hard. Let $\mathbb{D}_{\text{coar}} = \{D_{\text{coar}}^{(m)}\}_{m=1}^{M}$ be a **diverse collection** of document types with a **coarse-grained** label set $\mathbb{C}_{\text{coar}}$, and let D_{fine} be the **specific complex** domain with a **fine-grained** label set $\mathbb{C}_{\text{fine}}$. We first perform Stage 1 on the diverse data of OmniLayout-1M to acquire basic layout abilities in spatial organization, and then conduct Stage 2 on a fine-grained annotated dataset (*e.g.*, M^{6}Doc) to adapt to the target domain with complex element categories as shown in Fig. 3.

$$\underbrace{\left(\mathbb{D}_{\text{coar}},\ \mathbb{C}_{\text{coar}}\right)}_{\substack{\text{Stage 1: Coarse Learning [EASY]}\\ \text{(Diverse Domains, Coarse-grained Labels)}}} \xrightarrow[\text{Transfer}]{} \underbrace{\left(D_{\text{fine}},\ \mathbb{C}_{\text{fine}}\right)}_{\substack{\text{Stage 2: Fine Adaptation [HARD]}\\ \text{(Specific Domain, Fine-grained Labels)}}} \tag{2}$$

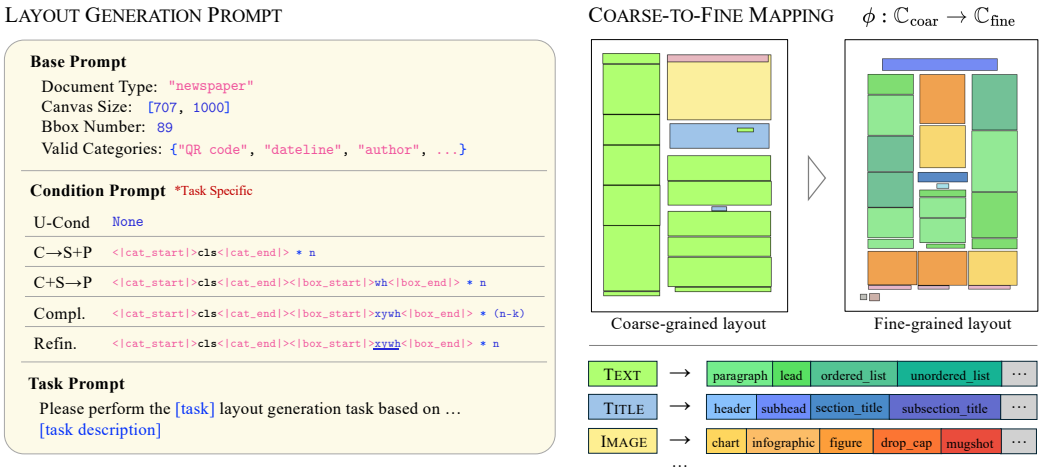


Figure 3: **Overview of Our Layout Generation Framework.** Unified layout generation prompt (base metadata + task-specific conditions for U-Cond, C→S+P, C+S→P, Completion, Refinement) and a Coarse-to-Fine mapping ϕ that transfers priors from diverse coarse labels to domain-specific fine categories.

3.2 LAYOUT GENERATION MODELING

We cast layout generation as conditional sequence modeling over a unified token space that encodes both semantic categories and normalized bounding boxes. Given a layout $\mathcal{L}$ serialized into a sequence of discrete tokens $T = (t_1, t_2, \ldots, t_K)$, the model is trained to maximize the conditional log-likelihood of this sequence.

Layout Generation Tasks. We follow the task setting introduced in (Zhang et al., 2025), and use five conditioning regimes that factor layout generation into category (C), size (S), and position (P), enabling controllable synthesis, constrained placement, completion, and editing: (1) **U-Cond**: Unconditional generation without external constraints. (2) **C→S+P**: Given the category of each element, the model predicts both its size and position. (3) **C+S→P**: The position of each element is masked; the model infers it from the provided category and size. (4) **Completion**: A subset (0–20%) of elements is retained on the page; the model completes the remaining layout to form a coherent structure. (5) **Refinement**: Geometric attributes are perturbed by Gaussian noise $\mathcal{N}(0, 10^{-2})$; the model recovers the original layout.

Layout Representation & Generation Prompt. Each element $e_i = (c, x, y, w, h)$ is serialized with a prefix-aware encoding <|cat_start|> c <|cat_end|><|box_start|> $0x$ $1y$ $2w$ $3h$ <|box_end|>, where coordinates x, y, w, h are normalized and uniformly quantized to $[0, 999]$. This unified serialization enables partial tuples for conditioning. The page-level *layout generation prompt* concatenates (i) a base header (document type, canvas size, bbox count, valid categories) and (ii) a task-specific *condition list* that supplies none / categories only / categories+wh / partial or perturbed tuples for, respectively, U-Cond, C→S+P, C+S→P, Completion, and Refinement, as shown in Fig. 3.

3.3 COARSE-TO-FINE LEARNING

The above formulation defines layout representation and conditioning regimes. A key challenge is how to train models that generalize across diverse document types and complex element categories. Directly learning fine-grained structures from limited data risks overfitting and poor transfer. We therefore adopt a *Coarse-to-Fine learning paradigm*, where the model first acquires robust spatial priors and structural regularities from diverse domains with coarse-grained labels, and then adapts to specific domains with fine-grained supervision. This staged strategy allows the model to progress from easy to hard, aligning training objectives with increasing complexity.

Coarse-grained Learning. The coarse-grained pre-training stage aims to establish a strong foundation for layout generation by harnessing the diversity of pre-training domains $\mathbb{D}_{\text{coar}}$. At this stage, the model is exposed to a wide range of document types, enabling it to acquire a broad understanding of document structures and the spatial relationships among various layout elements. Central to our approach is the unified representation of layout elements and the harmonization of label spaces. To

Task	Method	Textbook				Newspaper				Magazine				Exam				Academic			
		FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑
U-Cond	LayoutDM	180.25	0.024	0.310	-	281.56	0.008	0.628	-	281.91	0.233	0.462	-	287.58	0.131	0.382	-	153.66	0.440	0.362	-
	LACE	251.41	0.001	3.206	-	423.21	0.001	4.982	-	325.67	0.001	6.789	-	325.45	0.002	3.602	-	276.05	0.001	9.980	-
	LayoutPrompter	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	LGGPT	197.81	0.980	1.049	0.000	154.20	2.591	0.350	0.000	162.94	3.190	0.813	0.038	157.11	1.324	0.135	0.047	236.72	0.533	**0.100**	0.000
	Ours	**40.28**	**0.219**	**0.102**	**0.288**	**39.73**	**0.015**	**0.084**	0.000	**41.82**	**0.089**	**0.151**	**0.266**	**40.32**	**0.072**	**0.182**	**0.236**	**36.48**	**0.089**	0.062	**0.415**
C→S+P	LayoutDM	178.24	0.520	0.246	0.041	288.98	0.010	0.582	0.091	271.52	0.401	0.453	0.081	296.86	0.171	0.365	0.043	141.28	1.458	0.422	0.068
	LACE	187.31	0.005	2.345	0.025	308.24	0.009	2.873	0.000	220.00	0.001	0.862	0.069	276.12	0.010	0.442	0.009	212.41	0.006	2.645	0.088
	LayoutPrompter	44.67	0.512	0.191	**0.166**	124.45	0.312	0.899	0.160	65.24	0.899	0.362	**0.224**	46.56	0.262	0.439	0.098	20.89	0.221	0.264	0.216
	LGGPT	177.91	0.990	0.463	0.064	167.39	2.731	0.444	0.000	172.45	3.161	1.133	0.026	186.38	1.392	0.229	0.032	244.44	0.710	3.182	0.000
	Ours	**18.38**	**0.228**	**0.121**	0.154	**10.71**	**0.014**	**0.086**	**0.185**	**21.08**	**0.092**	**0.138**	0.221	**8.68**	**0.074**	**0.241**	**0.121**	**16.84**	**0.084**	**0.070**	**0.246**
C+S→P	LayoutDM	174.82	0.471	0.452	0.093	285.43	**0.010**	0.679	0.135	172.01	0.441	0.537	0.136	144.29	0.162	0.468	0.066	76.72	1.135	0.479	0.118
	LACE	28.79	0.001	6.345	0.015	256.08	0.005	4.158	0.006	196.78	0.002	6.015	0.048	160.28	0.008	6.327	0.050	99.86	0.008	1.402	0.097
	LayoutPrompter	42.38	**0.224**	0.469	0.199	126.78	0.219	1.387	0.156	41.52	0.245	0.442	0.235	14.58	**0.042**	0.341	0.138	14.58	0.203	0.332	0.286
	LGGPT	181.61	0.940	0.587	0.000	185.72	2.930	0.402	0.000	169.95	3.297	1.102	0.038	180.76	1.441	0.189	0.026	244.67	0.561	3.151	0.000
	Ours	**16.92**	0.366	**0.122**	**0.219**	**6.13**	0.021	**0.188**	**0.240**	**20.74**	**0.130**	**0.174**	**0.256**	**5.42**	0.083	**0.235**	**0.200**	**9.02**	**0.162**	**0.085**	**0.360**
Compl.	LayoutDM	172.35	0.012	0.429	0.000	270.15	0.007	0.704	0.000	260.15	0.113	0.557	0.000	255.17	0.073	0.459	0.000	134.51	0.370	0.418	0.000
	LACE	268.36	0.185	0.158	0.000	432.76	0.034	2.865	0.000	316.32	0.043	0.342	0.000	332.15	0.071	**0.218**	0.000	284.16	0.107	0.768	0.000
	LayoutPrompter	46.76	0.491	0.244	0.169	86.99	0.357	0.481	0.000	39.02	0.676	0.234	**0.342**	32.83	**0.066**	0.321	0.109	32.24	0.168	0.287	**0.642**
	LGGPT	192.32	1.180	0.892	0.158	160.25	2.696	0.335	0.000	153.43	3.511	0.743	0.052	153.79	1.461	0.167	0.000	242.17	0.975	2.812	0.000
	Ours	**31.58**	**0.235**	**0.123**	**0.478**	**22.48**	**0.013**	**0.098**	0.000	**38.56**	**0.098**	**0.153**	0.288	**25.92**	0.068	0.203	**0.310**	**30.56**	**0.106**	**0.070**	0.620
Refin.	LayoutDM	124.85	0.521	0.269	0.123	264.69	0.010	0.624	0.142	203.91	0.361	0.431	0.163	228.59	0.138	0.410	0.090	63.15	1.212	0.411	0.132
	LACE	143.95	0.174	0.736	0.165	291.35	**0.013**	2.047	0.234	216.23	0.141	0.693	0.256	214.19	**0.048**	0.921	0.187	42.89	0.219	0.621	0.254
	LayoutPrompter	11.82	0.149	0.511	0.672	113.84	0.154	1.216	0.647	10.26	0.919	0.430	0.689	10.87	0.072	0.422	0.502	9.23	0.177	0.437	0.667
	LGGPT	217.05	1.031	1.282	0.010	220.33	3.675	0.695	0.000	145.76	3.665	0.662	0.024	215.24	2.420	0.578	0.000	246.62	0.408	3.342	0.010
	Ours	**4.51**	**0.317**	**0.145**	**0.681**	**10.60**	0.017	**0.064**	**0.732**	**4.73**	**0.113**	**0.072**	**0.752**	**6.66**	0.079	**0.295**	**0.641**	**8.25**	**0.132**	**0.138**	**0.708**
Test Data		-	0.289	0.010	-	-	0.012	0.051	-	-	0.083	0.054	-	-	0.062	0.266	-	-	0.097	0.126	

Table 2: **Comparison with Layout Generation Experts across Five Document Types in M^{6}Doc.** For metrics, Ali. and Ove. denote *Alignment* and *Overlap*, → means closer to ground truth is better. For tasks, Compl. and Refin. denote Completion and Refinement, respectively. "-" indicates not applicable.

promote generalization across domains, we employ a coarse-grained label set $\mathbb{C}_{\text{coar}}$ that covers essential document components, such as text, table, image, and title, as well as associated classes like caption and footnote. This unified labeling strategy ensures the model learns transferable structural priors applicable to diverse document layouts.

Fine-grained Adaptation. Given a target domain D_{fine} with fine-grained labels $\mathbb{C}_{\text{fine}}$, we adapt the foundation model under a supervised sequence-modeling objective. The adaptation relies on a label mapping $\phi : \mathbb{C}_{\text{coar}} \to \mathbb{C}_{\text{fine}}$, where each coarse class is expanded into its fine-grained descendants (*e.g.*, "text" $\mapsto$ {"paragraph", "lead", "ordered_list"}). We fine-tune models for heterogeneous targets containing multiple document types (*e.g.*, NEWSPAPER, EXAM, ACADEMIC), yielding sharper, type-aware categories while preserving the general structural priors and cross-type generalization acquired during coarse-grained pretraining.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We first conduct coarse learning on OmniLayout-1M to acquire general layout patterns, and then perform fine-grained adaptation on five document types from M^{6}Doc (Cheng et al., 2023) datase, which includes: (1) **Textbook**: 2,080 samples spanning three grade levels and nine subjects, annotated with 42 element categories. (2) **Newspaper**: 1,000 samples from People's Daily[1] and The Wall Street Journal[2], with 42 element categories. (3) **Magazine**: 2,000 samples from globally recognized publishers such as Time USA, annotated with 26 categories. (4) **Exam**: 2,000 exam paper samples covering the same nine subjects as textbooks, with 31 categories. (5) **Academic**: 1,000 samples sourced from the arXiv, with 25 categories. We follow its original 6:1:3 split for training, validation, and testing.

Evaluation Metrics. Following previous work, we conduct our experiments using four standard metrics, grouped into two categories. (1) **Similarity Assessment**: *Fréchet Inception Distance (FID)* (Heusel et al., 2017) measures the similarity between generated and ground truth layouts by comparing their feature distributions in the embedding space of a deep neural network trained on corresponding images; *Maximum Intersection over Union (mIoU)* (Kikuchi et al., 2021) evaluates spatial alignment by optimally matching generated layouts with their ground truth counterparts to maximize the average IoU. (2) **Aesthetic Consistency**: We also adopt the *Alignment* (scaled by 100× for better visualization) and *Overlap* metrics from LayoutGAN++ (Kikuchi et al., 2021) to evaluate generation quality from the perspective of aesthetic principles.

[1] https://en.people.cn/

[2] https://www.wsj.com/

Task	Method	Textbook				Newspaper				Magazine				Exam				Academic			
		FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑
U-Cond	GPT-4o	135.32	0.017	**0.007**	0.060	193.13	0.020	0.007	0.000	236.11	0.015	**0.040**	0.089	163.94	0.020	0.010	0.000	135.60	0.006	0.006	0.000
	Gemini-2.5*	147.88	**0.264**	6.490	0.154	194.77	0.078	0.098	0.000	118.78	0.041	0.213	0.226	140.48	**0.071**	**0.313**	0.000	57.36	0.347	0.089	0.000
	Claude-3.7*	96.23	0.102	0.145	0.236	171.01	0.079	**0.031**	0.000	165.76	0.030	0.182	0.000	114.90	0.014	0.038	0.000	106.98	0.030	**0.101**	0.000
	Ours	**40.28**	0.219	0.102	**0.288**	**39.73**	**0.015**	0.084	0.000	**41.82**	**0.089**	0.151	**0.266**	**40.32**	0.072	0.182	**0.236**	**36.48**	**0.089**	0.062	**0.415**
C→S+P	GPT-4o	103.15	0.084	0.072	0.119	202.84	0.002	0.112	0.028	165.36	0.055	0.164	0.097	107.17	0.040	0.049	0.082	90.67	0.224	0.027	0.123
	Gemini-2.5*	54.74	0.175	**0.060**	0.078	69.40	0.569	0.666	0.070	53.32	0.065	**0.104**	0.101	42.25	**0.053**	0.063	0.036	31.79	0.351	0.051	0.084
	Claude-3.7*	42.99	0.117	0.068	0.127	53.62	0.001	0.520	0.079	87.00	0.071	0.109	0.126	27.96	0.041	0.080	0.087	66.22	0.138	**0.075**	0.139
	Ours	**18.38**	**0.228**	0.121	**0.154**	**10.71**	**0.014**	**0.086**	**0.185**	**21.08**	**0.092**	0.138	**0.221**	**8.68**	0.074	**0.241**	**0.121**	**16.84**	**0.084**	0.070	**0.246**
C+S→P	GPT-4o	64.67	0.448	0.363	0.091	106.97	0.043	4.759	0.052	112.38	0.332	0.765	0.076	61.67	0.187	0.905	0.049	58.49	0.743	0.852	0.075
	Gemini-2.5*	139.01	1.103	0.751	0.057	117.93	0.034	6.159	0.039	110.78	0.259	0.969	0.085	43.38	0.138	0.937	0.050	62.75	0.994	0.788	0.063
	Claude-3.7*	26.86	0.147	**0.103**	0.136	30.80	0.002	0.300	0.127	39.05	**0.086**	0.247	0.160	12.69	**0.054**	0.170	0.096	26.47	0.236	**0.116**	0.161
	Ours	**16.92**	**0.366**	0.122	**0.219**	**6.13**	**0.021**	**0.188**	**0.240**	**20.74**	0.130	**0.174**	**0.256**	**5.42**	0.083	**0.235**	**0.200**	**9.02**	**0.162**	0.085	**0.360**
Compl.	GPT-4o	61.20	**0.240**	**0.051**	**0.522**	97.60	0.227	**0.057**	0.000	155.36	0.115	**0.072**	0.075	116.18	0.124	0.068	0.000	93.49	0.068	0.063	0.000
	Gemini-2.5*	108.60	0.511	7.337	0.219	95.02	0.165	0.252	0.000	111.59	0.209	0.463	**0.355**	91.24	0.225	0.929	0.210	52.29	0.254	0.778	0.284
	Claude-3.7*	61.14	0.135	0.054	0.275	90.96	0.025	0.072	0.000	118.13	0.062	0.103	0.195	63.31	**0.063**	0.042	0.000	77.85	0.067	0.053	0.000
	Ours	**31.58**	0.235	0.123	0.478	**22.48**	**0.013**	0.098	0.000	**38.56**	**0.098**	0.153	0.288	**25.92**	0.068	**0.203**	**0.310**	**30.56**	**0.106**	**0.070**	**0.620**
Refin.	GPT-4o	12.71	0.371	0.162	0.616	67.25	0.040	0.172	0.628	7.76	0.198	0.108	0.654	**5.88**	0.121	**0.278**	0.577	3.27	0.178	0.127	0.618
	Gemini-2.5*	23.88	0.394	0.171	0.631	8.92	0.034	0.186	0.627	20.76	0.206	0.111	0.661	10.59	0.125	0.350	0.585	5.78	0.203	0.167	0.624
	Claude-3.7*	15.02	**0.272**	0.176	0.603	**3.86**	0.028	0.118	0.635	17.93	0.116	0.304	0.635	6.08	0.095	0.375	0.584	**1.67**	0.136	**0.127**	0.651
	Ours	**4.51**	0.317	**0.145**	**0.681**	10.60	**0.017**	**0.064**	**0.732**	**4.73**	**0.113**	**0.072**	**0.752**	6.66	**0.079**	0.295	**0.641**	8.25	**0.132**	0.138	**0.708**
Test Data		-	0.289	0.010	-	-	0.012	0.051	-	-	0.083	0.054	-	-	0.062	0.266	-	-	0.097	0.126	

Table 3: **Comparison with Powerful General-purpose LLMs in 0-shot Setting across Five Document Types in M^{6}Doc.** For models, Gemini-2.5* and Claude-3.7* denote Gemini-2.5-Flash and Claude-3.7-Sonnet.

Implementation Details. We choose Qwen2.5-0.5B-Instruct[3] as our base model. In coarse-grained learning stage, we constructed 9M samples from OmniLayout-1M across five tasks with a ratio of 1:1:1:3:3. Our model was then trained for 1 epoch on 40 NVIDIA A100 GPUs with a batch size of 16 per device and an initial learning rate of 1e-4, which took about 20 hours. For subsequent fine-grained adaptation stage, we adopted the same data construction strategy and trained for 5 epochs on different categories, respectively. In general, this process was conducted using 8 NVIDIA A100 GPUs, taking about 2 hours, with a batch size of 16 per device and an initial learning rate of 5e-5.

4.2 Comparison with Layout Experts

Baselines. For layout generation experts, we compare our approach against four representative methods spanning two major categories: (1) **Diffusion-based Models**: LayoutDM (Inoue et al., 2023) and LACE (Chen et al., 2024). (2) **LLM-based Methods**: LayoutPrompter (Lin et al., 2023) and LGGPT (Zhang et al., 2025). Several work are excluded from comparison for the following reasons: (1) **Early Vintage.** Earlier research such as LayoutGAN++ (Kikuchi et al., 2021) and LayoutFormer++ (Jiang et al., 2023) are no longer suitable as fair baselines against modern models. (2) **Unavailable or Buggy Implementation.** The latest work like LayoutCoT (Shi et al., 2025) or LayoutRAG (Wu et al., 2025) lack publicly available code repositories, and the released implementation of LayoutNUWA (Tang et al., 2023) is hard to reproduce. (3) **Poor Convergence.** We trained on LayoutFlow (Guerreiro et al., 2024) for more than 100K epochs but failed to converge to a satisfactory result.

Analysis. We adopt a unified domain-specific training strategy, selecting the checkpoint with the lowest validation loss for fair comparison. The detailed results are shown in Table 2. We observe that: (1) For the two diffusion-based models, the overall performance on *FID* is unsatisfactory. This can be attributed to the intrinsic nature of diffusion models, which require substantially more training data and longer convergence time to accurately learn probability distributions. As a result, they fail in low-resource and complex domains. Although LACE achieves significant improvement in element *alignment* through post-processing, it still struggles to control *overlap*. (2) For the three LLM-based models, thanks to the autoregressive formulation and strong long-context modeling capability, they can naturally follow aesthetic layout rules without the need for specially designed post-processing. An exception is LGGPT, where the underlying GPT2-XL (Radford et al., 2019) often produces incoherent or nonsensical outputs when handling long prompt sequences, a problem not observed on shorter-sequence datasets like PubLayNet. Compared to these baselines, our model achieves consistently superior results across all metrics, with particularly notable gains on *mIoU*.

4.3 Comparison with General-purpose LLMs

Baselines. For general-purpose LLMs, we select three powerful LLMs: GPT-4o (OpenAI, 2024), Gemini-2.5-Flash (Google, 2025a), and Claude-3.7-Sonnet (Anthropic, 2025), chosen for their

[3] https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct

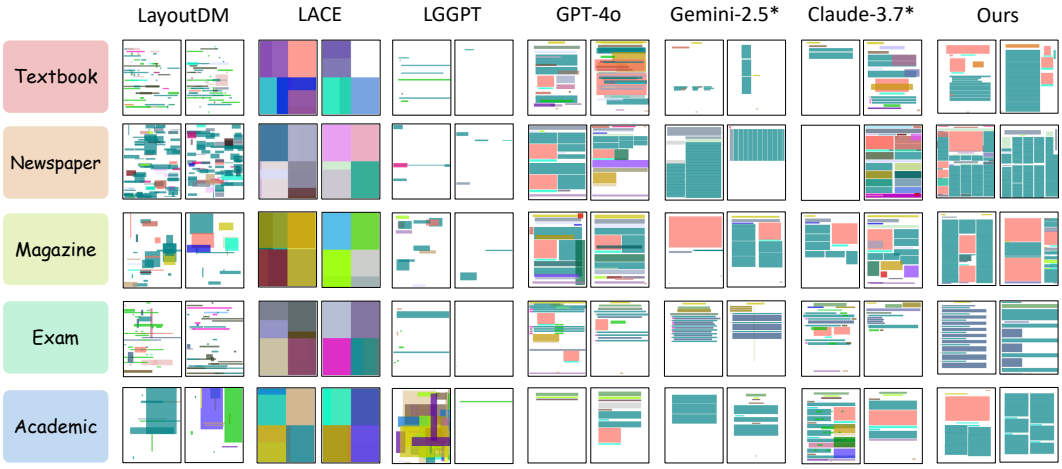


Figure 4: **Visualization Examples of Various Methods with U-Cond Task.** For general-purpose LLMs, we adopt the strongest 5-shot setting.

strong long-context capability, with the latter two also offering advanced reasoning abilities relevant to complex layout generation.

Analysis. We conduct evaluation under the classic 0/1/5-shot settings to assess whether existing LLMs can achieve competitive performance on conditional document layout generation solely through in-context learning. Owing to the page limit, Table 3 reports zero-shot results, while few-shot results are provided in Appendix Table 5. We observe that: (1) Under zero-shot setting, all general-purpose LLMs achieve reasonably good *alignment* and *overlap*, but exhibit high stochasticity, sometimes leading to extreme outliers. For example, Gemini-2.5-Flash attains an unexpectedly high *overlap* score on the textbook while performing unconditional task. Moreover, complex layouts such as Newspaper remain the most challenging, as reflected by the highest average *FID*, whereas performance on Academic is relatively better, likely because such formats are more prevalent in pre-training corpora. Among the three models, Claude-3.7-Sonnet delivers the best results. (2) Under few-shot settings, all LLMs improve as the number of shots increases, confirming that in-context learning indeed enables better layout generation. Nevertheless, although additional shots yield better performance, the improvement tends to converge to an intrinsic upper bound, and comes at the cost of longer input sequences, higher API expenses, and slower inference. The visualization results of different methods on the U-Cond task are shown in Fig. 4, and more examples of our OmniLayout-LLM can be found in Appendix C.

4.4 Ablation Study

Task	Param	FID↓	Ali.→	Ove.→	mIoU↑
U-Cond	3B	**35.81**	0.009	**0.052**	0.000
	1.5B	36.58	**0.013**	0.105	0.000
	0.5B	39.73	0.015	0.084	0.000
C→S+P	3B	26.88	0.007	0.110	0.000
	1.5B	**10.63**	**0.012**	0.097	0.179
	0.5B	10.71	0.014	**0.086**	**0.185**
C+S→P	3B	17.12	**0.020**	**0.181**	**0.320**
	1.5B	**5.65**	0.022	0.205	0.238
	0.5B	6.13	0.021	0.188	0.240
Compl.	3B	27.08	0.008	0.125	0.000
	1.5B	26.86	0.011	**0.093**	0.000
	0.5B	**22.48**	**0.013**	0.098	0.000
Refin.	3B	67.24	0.017	0.062	0.725
	1.5B	**6.98**	0.017	**0.061**	0.730
	0.5B	10.60	**0.017**	0.064	**0.732**
Test Data		-	0.012	0.051	-

Task	Stage	FID↓	Ali.→	Ove.→	mIoU↑
U-Cond	F.	42.98	0.017	8.308	0.000
	C.	249.1	0.016	0.388	0.000
	Both	**39.73**	**0.015**	**0.084**	0.000
C→S+P	F.	14.88	0.024	0.493	0.164
	C.	246.4	0.016	0.357	0.000
	Both	**10.71**	**0.014**	**0.086**	**0.185**
C+S→P	F.	11.24	0.023	0.476	0.220
	C.	235.8	0.021	1.162	0.000
	Both	**6.13**	**0.021**	**0.188**	**0.240**
Compl.	F.	36.99	0.015	6.627	0.000
	C.	248.79	0.015	0.480	0.000
	Both	**22.48**	**0.013**	0.098	0.000
Refin.	F.	22.07	0.023	1.452	0.618
	C.	254.98	0.018	0.386	0.000
	Both	**10.60**	**0.017**	**0.064**	**0.732**
Test Data		-	0.012	0.051	-

Table 4: **Ablation on Model Sizes and Learning Stages.** F. and C. denote Fine-grained Adaptation and Coarse-grained Learning only, respectively.

In this section, we perform ablation studies on the number of parameters and the two stages of the Coarse-to-Fine framework, conducted in the most challenging newspaper domain. The results are shown in Table 4. We observe that: (1) For model size, the overall differences are marginal. The 3B model achieves slightly lower *Alignment* scores, but its *FID* increases in most tasks. For instance, in the C+S→P task the 3B model yields an *FID* that is 2.79× higher than that of the 0.5B model. This phenomenon is likely due to the inherent volatility of *FID* when evaluated on limited test samples. (2) For *Coarse-to-Fine learning paradigm*, coarse-grained learning brings substantial gains in orga-

nizing and perceiving the overall layout, as evidenced by a sharp reduction in *Overlap*. Fine-grained adaptation further enhances the model's ability to output diverse and detailed element labels across document types, enabling more sophisticated layout generation. Notably, multiple zero scores in *mIoU* are attributed to its definition: it requires exact label-level matches, which are often absent in complex multi-element layout generation, particularly in U-Cond and Completion tasks.

5 Related Work

Document Layout Dataset. Existing layout datasets largely stem from parsing tasks and initially focused on academic articles. PubLayNet (Zhong et al., 2019) auto-annotates papers from PubMed Central via XML matching, while DocBank (Li et al., 2020) uses weak supervision on arXiv. Recent efforts broaden document types but remain small and often require manual annotation. Although DocLayNet (Pfitzmann et al., 2022) and D^{4}LA (Da et al., 2023) cover six and twelve types, respectively, they still contain only on the order of 10^4 pages. M^{6}Doc (Cheng et al., 2023) and OmniDocBench (Ouyang et al., 2025) reflect real-world formats (e.g., textbooks, newspapers) but are even smaller, and the latter serves purely as a benchmark without a training split.

Layout Generation. Document layout generation has gained traction. Early methods used GANs or transformers: LayoutGAN++ (Kikuchi et al., 2021) enhances the GAN framework with transformer blocks and optimizes latent codes to achieve constrained layout generation. LayoutTransformer (Gupta et al., 2021) leverages self-attention to learn contextual relationships among layout elements, BLT (Kong et al., 2022) adopts a non-autoregressive bidirectional transformer that iteratively refines layouts by masking and predicting low-confidence attributes through a hierarchical sampling strategy. LayoutFormer++ (Jiang et al., 2023) employs constraint tokenization and restricted decoding space to strike a balance between user constraint satisfaction and overall layout quality. More recently, diffusion-based methods have gained attention. LayoutDM (Inoue et al., 2023) and LACE (Chen et al., 2024)) denoise element coordinates and labels in discrete/continuous spaces, respectively, and attempt to inject hard/soft constraints. Flow-based LayoutFlow (Guerreiro et al., 2024) frames the task as flow matching, speeding training and inference.

Large Language Model. With the remarkable success of LLMs in sequence generation tasks (OpenAI, 2024; Anthropic, 2025; Google, 2025b), autoregressive generation has become the mainstream paradigm for document layout generation in recent years. LayoutPrompter (Lin et al., 2023) casts layouts into unified HTML representations and employs adaptive exemplar retrieval to enable in-context learning. LayoutCoT (Shi et al., 2025) leverages the deep reasoning capabilities of general-purpose LLMs through chain-of-thought prompting, substantially improving the performance and practicality of training-free layout generation. LayoutRAG (Wu et al., 2025) retrieves optimal reference layouts from a layout database and introduces a condition-modulated attention module to selectively incorporate prior knowledge. While effective, these approaches remain largely domain-specific and rely heavily on prompt engineering or retrieval heuristics. In contrast, LayoutNUWA Tang et al. (2023) and LGGPT (Zhang et al., 2025) pioneer domain-agnostic paradigms that fully exploit the generalization strength of LLMs: the former formulates the task as HTML code completion, whereas the latter demonstrates that pure string-based input–output reduces redundant tokens and yields better efficiency.

6 Conclusion

In this work, we move beyond the domain limitations of previous studies and explore complex document layout generation with LLMs. To address the scarcity of diverse training data, we introduce OmniLayout-1M, the first million-scale dataset for document layouts, covering six common types such as newspapers and textbooks. Moreover, leveraging the strong capability of LLMs in long-sequence generation, we propose a *Coarse-to-Fine learning paradigm*: first acquiring fundamental aesthetic layout rules from comprehensive document types, and then performing fine-grained adaptation on specific complex domain. Our approach significantly surpasses both existing layout generation experts and powerful general-purpose LLMs. However, our experiments also reveal challenges, such as the inadequacy of current metrics when evaluating complex layouts under limited samples. We will continue to investigate these issues to further advance the field of Document AI.

Ethics Statement

The development of OmniLayout-1M and the associated model was guided by a commitment to ethical research practice. We acknowledge that our dataset, despite its scale and diversity, may contain inherent biases from its sources, which could be reflected in the models trained on it. We encourage users to be aware of these potential biases. Furthermore, we recognize that layout generation models could be misused to create misleading or fraudulent documents. Our research is intended for positive applications, such as enhancing document synthesis, streamlining content creation workflows, and improving document understanding. We disavow any malicious use of our work and hope that by making our methods and dataset public, we can foster further research into responsible and ethical generative AI for documents.

Reproducibility Statement

We are committed to the full reproducibility of our work. The core concepts and methodology of our *Coarse-to-Fine learning paradigm* are detailed in Section 3. Our experimental setup, including the datasets used, evaluation protocols, and baseline models, are described in Section 4. Further details on the OmniLayout-1M dataset, including more data statistics and analysis, are provided in Appendix D. To facilitate the reproduction of our results and to encourage further research, we will release our codebase, the OmniLayout-1M dataset, and the pretrained model weights.

References

Anthropic. Claude-3.7-sonnet. https://www.anthropic.com/news/claude-3-7-sonnet, 2025.

Anthropic. Claude sonnet 4, 2025. URL https://www.anthropic.com/claude/sonnet. Accessed: 2025-05-23.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Rajiv Jain, Zhiqiang Xu, Ryan Rossi, and Changyou Chen. Towards aligned layout generation via diffusion model with aesthetic constraints. *arXiv preprint arXiv:2402.04754*, 2024.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. 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*, pp. 15138–15147, 2023.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. URL https://arxiv.org/abs/2507.05595.

Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision grid transformer for document layout analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19462–19472, 2023.

Google. gemini-2.5-flash. https://deepmind.google/models/gemini/flash/, 2025a.

Google. gemini-2.5-pro. https://deepmind.google/models/gemini/pro/, 2025b.

Julian Jorge Andrade Guerreiro, Naoto Inoue, Kento Masui, Mayu Otani, and Hideki Nakayama. Layoutflow: flow matching for layout generation. In *European Conference on Computer Vision*, pp. 56–72. Springer, 2024.

Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1004–1014, 2021.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pp. 991–995. IEEE, 2015.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10167–10176, 2023.

Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Huayu Deng, Zhongkai Wu, Vuksan Mijovic, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. Layoutformer++: Conditional graphic layout generation via constraint serialization and decoding space restriction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18403–18412, 2023.

Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Constrained graphic layout generation via latent optimization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 88–96, 2021.

Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. Blt: Bidirectional layout transformer for controllable layout generation. In *European Conference on Computer Vision*, pp. 474–490. Springer, 2022.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm, 2025. URL https://arxiv.org/abs/2506.05218.

Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. Layoutprompter: Awaken the design ability of large language models. *Advances in Neural Information Processing Systems*, 36:43852–43879, 2023.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24838–24848, 2025.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. 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.

Alec Radford, Jeffrey Wu, and et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019.

Hengyu Shi, Junhao Su, Huansheng Ning, Xiaoming Wei, and Jialin Gao. Layoutcot: Unleashing the deep reasoning potential of large language models for layout generation. *arXiv preprint arXiv:2504.10829*, 2025.

Zecheng Tang, Chenfei Wu, Juntao Li, and Nan Duan. Layoutnuwa: Revealing the hidden layout expertise of large language models. *arXiv preprint arXiv:2309.09506*, 2023.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. Mineru: An open-source solution for precise document content extraction, 2024. URL https://arxiv.org/abs/2409.18839.

Yuxuan Wu, Le Wang, Sanping Zhou, Mengnan Liu, Gang Hua, and Haoxiang Li. Layoutrag: Retrieval-augmented model for content-agnostic conditional layout generation. *arXiv preprint arXiv:2506.02697*, 2025.

Peirong Zhang, Jiaxin Zhang, Jiahuan Cao, Hongliang Li, and Lianwen Jin. Smaller but better: Unifying layout generation with smaller large language models. *International Journal of Computer Vision*, 133(7):3891–3917, 2025.

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. *arXiv preprint arXiv:2410.12628*, 2024.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pp. 1015–1022. IEEE, 2019.

APPENDIX

A LLM USAGE STATEMENT

We used AI tools (Large Language Models) at a **minimal** level and *only* for linguistic polishing (grammar, spelling, punctuation, and minor word-choice/style edits). The tools did **not** change the original meaning, nor did they introduce any new ideas, claims, content, data, code, figures, analyses, or results. All technical contributions and writing decisions were authored and verified by the authors, and no confidential or proprietary data were provided to AI services.

Task	Method	Setting	Textbook				Newspaper				Magazine				Exam				Academic			
			FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑
U-Cond	GPT-4o	0-shot	135.32	0.017	**0.007**	0.060	193.13	0.020	0.007	0.000	236.11	0.015	**0.040**	0.089	163.94	0.020	0.010	0.000	135.60	0.006	0.006	0.000
		1-shot	97.34	0.105	0.099	0.061	100.18	0.031	**0.052**	0.000	177.61	0.038	0.215	0.158	111.87	0.043	0.049	0.000	100.70	0.023	0.011	0.000
		5-shot	71.56	0.149	0.083	0.177	54.34	0.032	0.074	0.000	143.70	0.051	0.198	0.174	76.48	0.049	0.085	0.000	90.93	0.060	0.024	**0.460**
	Gemini-2.5*	0-shot	147.88	**0.264**	6.490	0.154	194.77	0.078	0.098	0.000	118.78	0.041	0.213	0.226	140.48	**0.071**	0.313	0.000	57.36	0.347	0.089	0.000
		1-shot	84.16	0.159	0.552	0.221	74.14	0.015	0.098	0.000	88.91	0.063	0.152	0.214	90.10	0.037	**0.299**	0.000	82.93	0.107	0.054	0.000
		5-shot	57.30	0.173	0.219	0.202	30.55	**0.014**	0.094	0.000	66.30	0.190	0.136	**0.289**	70.05	0.038	0.086	0.000	51.41	0.074	0.061	0.322
	Claude-3.7*	0-shot	96.23	0.102	0.145	0.236	171.01	0.079	0.031	0.000	165.76	0.030	0.182	0.000	114.90	0.014	0.038	0.000	106.98	0.030	0.101	0.000
		1-shot	87.87	0.096	0.164	0.171	73.78	0.003	0.512	0.000	141.53	0.023	0.251	0.000	72.29	0.025	0.120	0.000	100.70	0.049	**0.112**	0.000
		5-shot	59.73	0.093	0.114	0.091	**25.57**	0.007	0.245	0.000	84.23	0.044	0.093	0.165	50.83	0.043	0.131	0.000	75.50	0.042	0.098	0.064
	Ours	-	**40.28**	0.219	0.102	**0.288**	39.73	0.015	0.084	0.000	**41.82**	**0**.089	0.151	0.266	**40.32**	0.072	0.182	**0.236**	**36.48**	**0.089**	0.062	0.415
C→S+P	GPT-4o	0-shot	103.15	0.084	0.072	0.119	202.84	0.002	0.112	0.028	165.36	0.055	0.164	0.097	107.17	0.040	0.049	0.082	90.67	0.224	0.027	0.123
		1-shot	72.34	0.111	**0.059**	0.115	152.61	0.006	0.230	0.043	134.33	0.108	0.178	0.102	58.18	0.030	0.054	0.087	64.51	0.166	0.040	0.125
		5-shot	51.50	0.146	0.074	0.118	91.87	**0.012**	0.312	0.059	104.85	0.112	0.193	0.110	31.96	0.040	0.087	0.091	41.54	0.1694	0.030	0.142
	Gemini-2.5*	0-shot	54.74	0.175	0.060	0.078	69.40	0.569	0.666	0.070	53.32	0.065	0.104	0.101	42.25	0.053	0.063	0.036	31.79	0.351	0.051	0.084
		1-shot	42.26	0.217	0.114	0.089	37.62	0.005	0.750	0.073	53.72	0.057	0.240	0.110	33.10	0.059	0.169	0.047	28.47	0.201	0.036	0.118
		5-shot	35.33	0.152	0.106	0.098	17.09	0.010	0.353	0.095	43.72	0.071	0.326	0.128	21.22	0.069	0.136	0.060	18.49	0.103	0.036	0.148
	Claude-3.7*	0-shot	42.99	0.117	0.068	0.127	53.62	0.001	0.520	0.079	87.00	0.071	0.109	0.126	27.96	0.041	0.080	0.087	66.22	0.138	0.075	0.139
		1-shot	33.61	0.111	0.083	0.127	36.33	0.003	0.416	0.096	55.11	0.067	0.204	0.128	16.32	0.021	0.227	0.095	44.09	0.109	0.109	0.144
		5-shot	18.86	0.103	0.102	0.122	13.14	0.004	0.332	0.000	27.34	0.061	0.112	0.139	17.34	0.025	0.130	0.100	17.49	0.062	0.105	0.175
	Ours	-	**18.38**	**0.228**	0.121	**0.154**	**10.71**	0.014	0.086	0.185	21.08	0.092	0.138	0.221	8.68	0.074	0.241	0.121	16.84	0.084	0.070	0.246
C+S→P	GPT-4o	0-shot	64.67	0.448	0.363	0.091	106.97	0.043	4.759	0.052	112.38	0.332	0.765	0.076	61.67	0.187	0.905	0.049	58.49	0.743	0.852	0.075
		1-shot	52.66	0.447	0.094	0.132	52.16	0.027	0.396	0.100	69.15	0.242	0.292	0.143	19.53	0.163	0.085	0.102	32.53	0.669	0.115	0.150
		5-shot	46.00	0.514	0.113	0.136	30.81	0.034	0.516	0.108	63.56	0.259	0.298	0.146	13.33	0.169	0.121	0.116	26.73	0.632	0.101	0.176
	Gemini-2.5*	0-shot	139.01	1.103	0.751	0.057	117.93	0.034	6.159	0.039	110.78	0.259	0.969	0.085	43.38	0.138	0.937	0.050	62.75	0.994	0.788	0.063
		1-shot	94.61	0.480	0.398	0.103	65.27	0.015	1.470	0.089	88.37	0.264	0.600	0.124	15.94	0.088	0.546	0.082	28.41	0.333	0.379	0.154
		5-shot	73.67	0.534	0.316	0.117	35.17	0.035	0.700	0.127	66.91	0.305	0.426	0.134	13.06	0.116	0.398	0.094	33.06	0.391	0.319	0.177
	Claude-3.7*	0-shot	26.86	0.147	0.103	0.136	30.80	0.002	0.300	0.127	39.05	0.086	0.247	0.160	12.69	0.054	0.170	0.096	26.47	0.236	0.116	0.161
		1-shot	20.04	0.136	0.130	0.146	17.91	0.005	0.381	0.147	33.47	0.078	0.226	0.171	9.75	0.028	0.274	0.113	22.69	0.142	0.143	0.186
		5-shot	21.47	0.159	0.082	0.156	18.75	0.006	0.409	0.159	34.77	0.061	0.331	0.178	7.64	0.031	0.390	0.122	14.14	0.257	0.084	0.222
	Ours	-	16.92	0.366	0.122	0.219	6.13	0.021	0.188	0.240	20.74	0.130	0.174	0.256	5.42	0.083	0.235	0.200	9.02	0.162	0.085	0.360
Compl.	GPT-4o	0-shot	61.20	0.240	0.051	0.522	97.60	0.227	0.057	0.000	155.36	0.115	0.072	0.075	116.18	0.124	0.068	0.000	93.49	0.068	0.063	0.000
		1-shot	44.68	0.309	0.045	0.131	84.59	0.144	0.058	0.000	130.47	0.125	0.083	0.139	78.13	0.168	0.060	0.320	70.53	0.112	0.068	0.000
		5-shot	33.44	0.340	0.052	0.268	50.11	0.090	0.118	0.000	86.84	0.143	0.099	0.169	39.22	0.184	0.087	0.290	46.80	0.135	0.060	0.438
	Gemini-2.5*	0-shot	108.60	0.511	7.337	0.219	95.02	0.165	0.252	0.000	111.59	0.209	0.463	0.355	91.24	0.225	0.929	0.210	52.29	0.254	0.778	0.284
		1-shot	86.44	0.553	0.699	0.302	81.28	0.184	5.176	0.000	89.80	0.135	0.227	0.168	50.65	0.243	0.578	0.402	35.75	0.247	0.300	0.272
		5-shot	65.85	0.394	0.360	0.310	45.02	0.049	0.1655	0.000	68.69	0.127	0.223	0.180	36.62	0.289	0.220	0.025	28.30	0.164	0.073	0.440
	Claude-3.7*	0-shot	61.14	0.135	0.054	0.275	90.96	0.025	0.072	0.000	118.13	0.062	0.103	0.195	63.31	0.063	0.042	0.000	77.85	0.067	0.053	0.000
		1-shot	54.28	0.103	0.190	0.259	53.40	0.008	0.173	0.000	100.98	0.042	0.209	0.232	45.10	0.061	0.110	0.000	68.76	0.070	0.082	1.000
		5-shot	29.74	0.225	0.111	0.331	18.54	0.012	0.167	0.000	47.29	0.072	0.148	0.172	25.88	0.071	0.139	0.089	36.78	0.112	0.128	0.329
	Ours	-	31.58	0.235	0.123	0.478	22.48	0.013	0.098	0.000	38.56	0.098	0.153	0.288	25.92	0.068	0.203	0.310	30.56	0.106	0.070	0.620
Refin.	GPT-4o	0-shot	12.71	0.371	0.162	0.616	67.25	0.040	0.172	0.628	7.76	0.198	0.108	0.654	5.88	0.121	0.278	0.577	3.27	0.178	0.127	0.618
		1-shot	6.75	0.392	0.157	0.646	23.67	0.042	0.175	0.639	7.63	0.190	0.104	0.670	4.32	0.125	0.286	0.599	2.10	0.162	0.134	0.640
		5-shot	10.25	0.397	0.180	0.650	31.24	0.040	0.170	0.639	8.92	0.194	0.116	0.672	4.89	0.124	0.291	0.601	1.38	0.198	0.134	0.658
	Gemini-2.5*	0-shot	23.88	0.394	0.171	0.631	8.92	0.034	0.186	0.627	20.76	0.206	0.111	0.661	10.59	0.125	0.350	0.585	5.78	0.203	0.167	0.624
		1-shot	9.72	0.386	0.165	0.656	1.05	0.036	0.182	0.638	8.62	0.196	0.105	0.665	5.21	0.124	0.328	0.597	3.79	0.188	0.159	0.646
		5-shot	8.18	0.392	0.158	0.657	1.27	0.034	0.183	0.637	6.87	0.200	0.104	0.669	1.02	0.124	0.321	0.606	1.32	0.189	0.163	0.661
	Claude-3.7*	0-shot	15.02	0.272	0.176	0.603	3.86	0.028	0.118	0.635	17.93	0.116	0.304	0.635	6.08	0.095	0.375	0.584	1.67	0.136	0.127	0.651
		1-shot	18.97	0.231	0.543	0.638	3.12	0.028	0.102	0.643	11.01	0.099	0.624	0.663	16.28	0.073	0.729	0.608	4.22	0.111	0.410	0.650
		5-shot	11.23	0.284	0.378	0.642	3.85	0.027	0.165	0.638	3.33	0.175	0.173	0.676	5.84	0.093	0.545	0.596	3.72	0.158	0.325	0.631
	Ours	-	4.51	0.317	0.145	0.681	10.60	0.017	0.064	0.732	4.73	0.113	0.072	0.752	6.66	0.079	0.295	0.641	8.25	0.132	0.138	0.708
	Test Data		-	0.289	0.010	-	-	0.012	0.051	-	-	0.083	0.054	-	-	0.062	0.266	-	-	0.097	0.126	

Table 5: **Comparison with Powerful General-purpose LLMs in 0/1/5-shot Setting across Five Document Types in M^{6}Doc.**

B Few-shot Performance of General-purpose LLMs

Due to space limitations, the complete 0/1/5-shot comparison results are reported in Table 5 of the appendix. In particular, evaluation of complex document layouts requires significantly longer inference time and more than 10,000 USD in API costs owing to excessive sequence length.

C Qualitative Results of OmniLayout-LLM across Diverse Domains

In this section, we demonstrate the qualitative results generated by OmniLayout-LLM: Fig. 5 for textbook and newspaper, Fig. 6 for magazine and exam, and Fig. 7 for academic. The visualization results demonstrate that OmniLayout-LLM can generate reasonable and aesthetically pleasing layouts for a wide variety of document types. Furthermore, it effectively adheres to the requirements of different generation tasks and adapts well to various constraints, showcasing its ability to perform under diverse conditions and tasks.

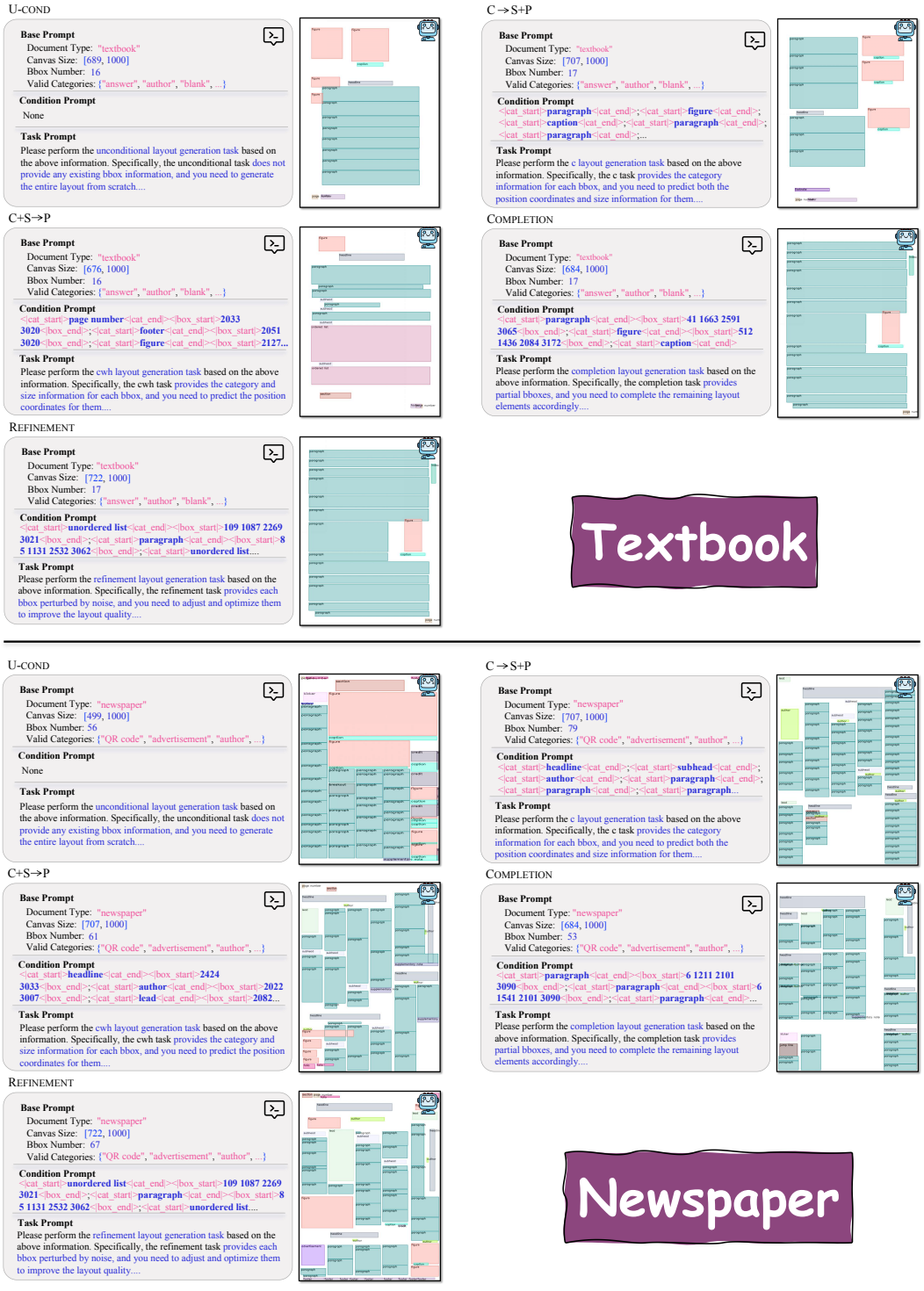


Figure 5: Visualization of Layouts Generated by OmniLayout-LLM (Textbook and Newspaper).

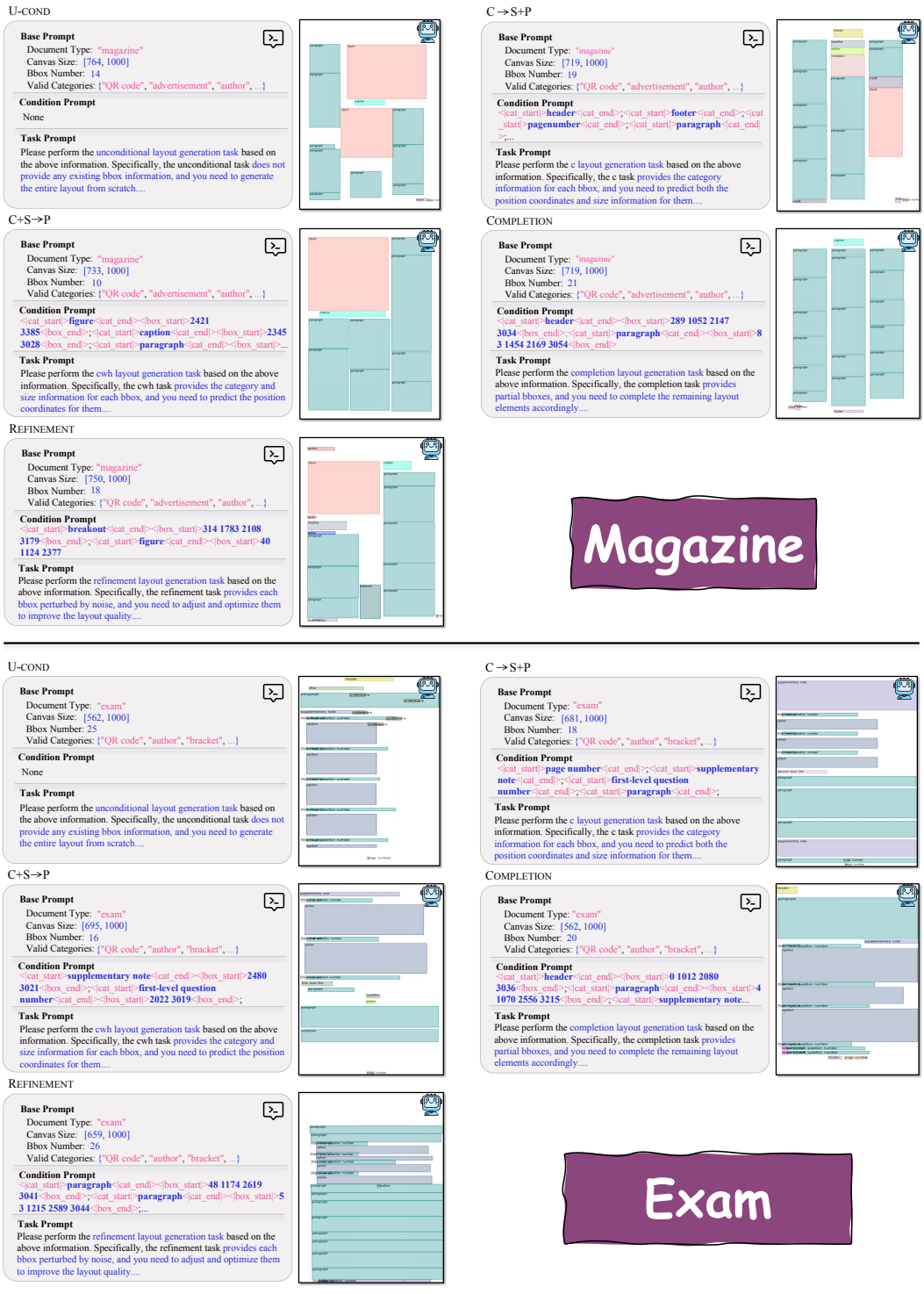


Figure 6: Visualization of Layouts Generated by OmniLayout-LLM (Magazine and Exam).

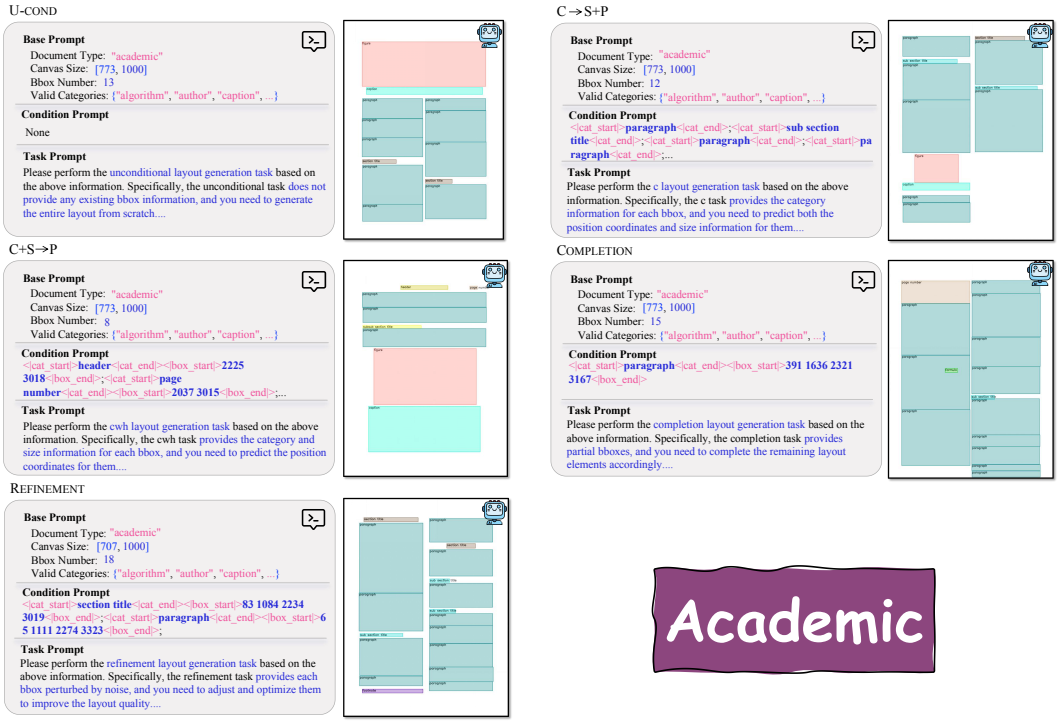
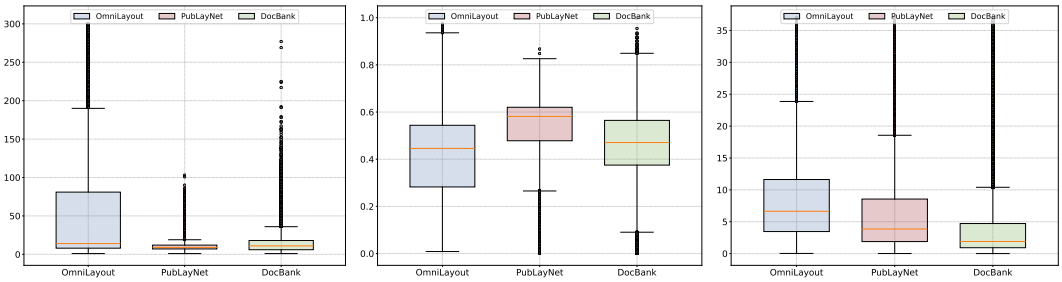


Figure 7: Visualization of Layouts Generated by OmniLayout-LLM (Academic).

D OMNILAYOUT-1M DATASET

D.1 ELEMENT-WISE STATISTICAL ANALYSIS

First, we analyze the diversity of OmniLayout-1M from the perspective of element distribution. Specifically, we examine element diversity in three aspects: the number of elements per page, the proportion of the layout area occupied by all elements on a page, and the aspect ratios of the elements. The data distribution is illustrated in Fig. 8. As can be observed, OmniLayout-1M exhibits significantly greater diversity in elements compared to PubLayNet and DocBank. This ensures the robustness of the pre-trained model, enabling our proposed method to adapt to various element types (with different aspect ratios and categories) and diverse layout attributes (with varying densities and numbers of elements) in downstream tasks.



(a) Element numbers per page. (b) Area ratio per page. (c) Element aspect ratio.

Figure 8: Element Statistical Analysis of OmniLayout-1M.

D.2 MORE VISUALIZATION EXAMPLES

In this section we present more visualization examples from our OmniLayout-1M dataset, accompanied by high-quality annotations extracted with MinerU (Wang et al., 2024). Visualization of

6 layout types: textbook (Fig. 9), newspaper (Fig. 10), magazine (Fig. 11), exam (Fig. 12), academic (Fig. 13), slide (Fig. 14) are shown.

D.3 LAYOUT DIVERSITY

Next, we visualize and compare the document layout diversity of PubLayNet, DocBank, and OmniLayout-1M as shown in Fig. 15 and Fig. 16. N indicates number of documents used for visualization. Compared with two-column format and Manhattan layout typical of academic papers in PubLayNet or DocBank, document layout in OmiLayout-1M significant variation and diversity.

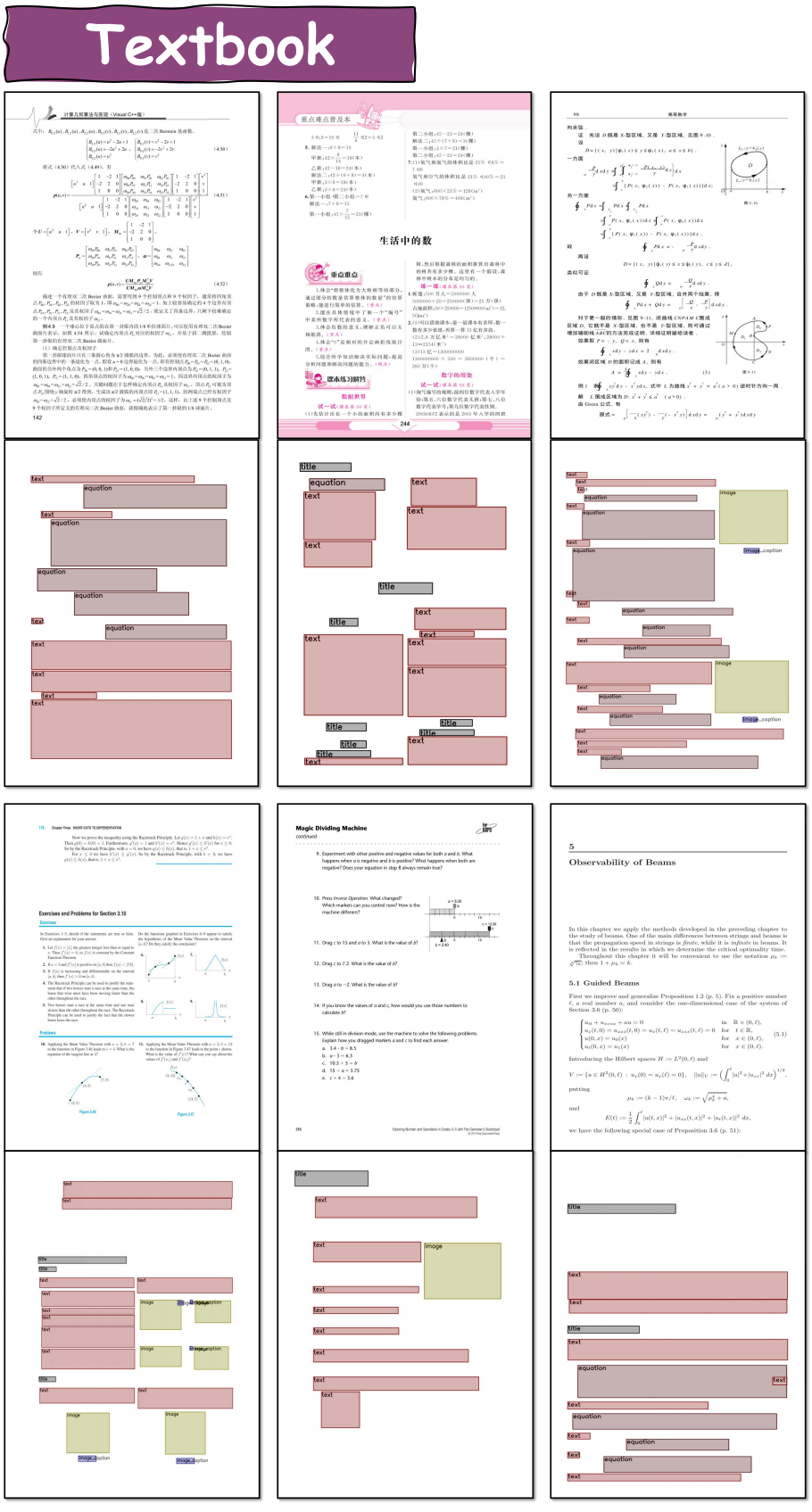


Figure 9: Visualization of Textbook Layout Data in OmniLayout-1M.

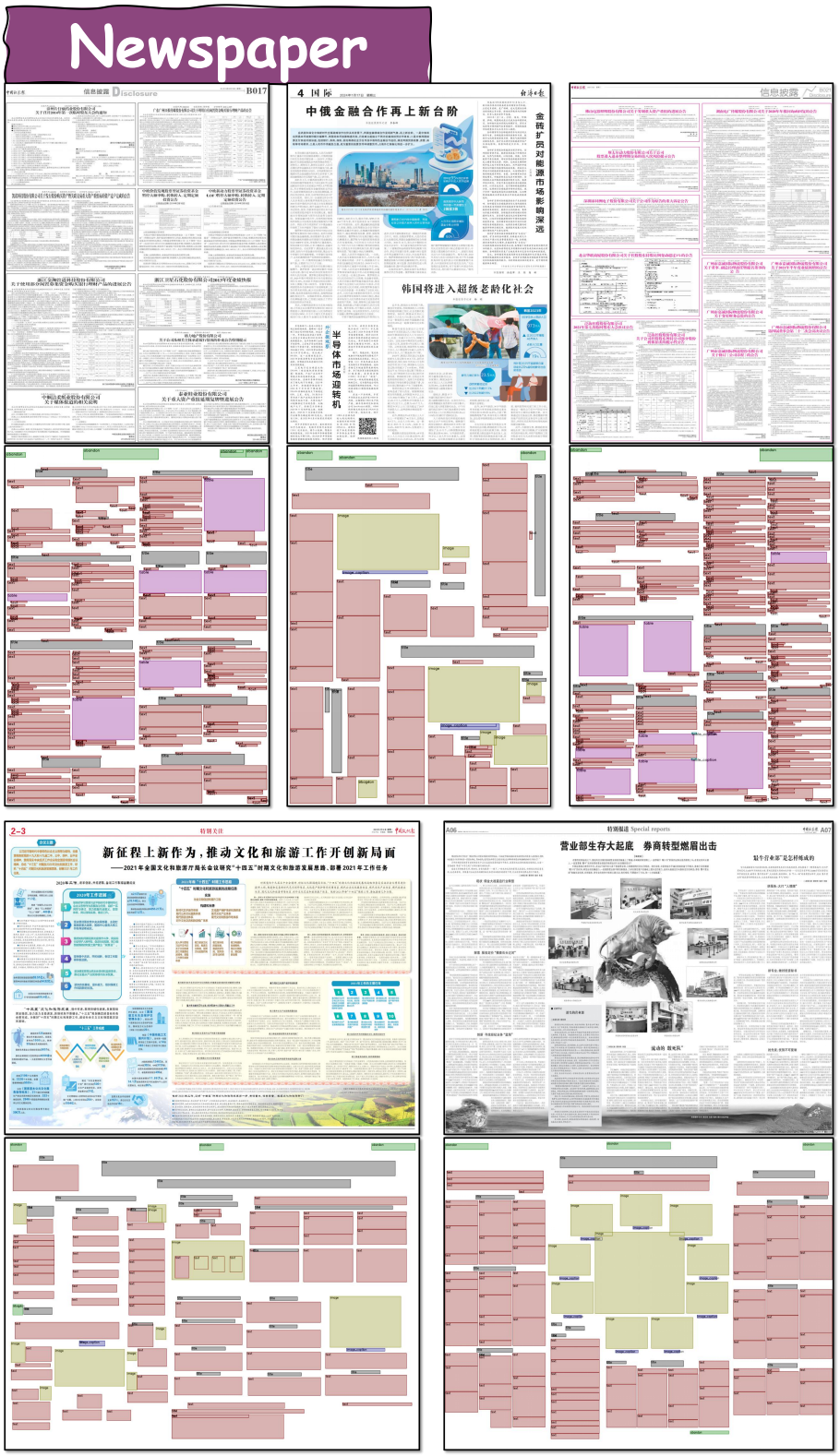


Figure 10: Visualization of Newspaper Layout Data in OmniLayout-1M.

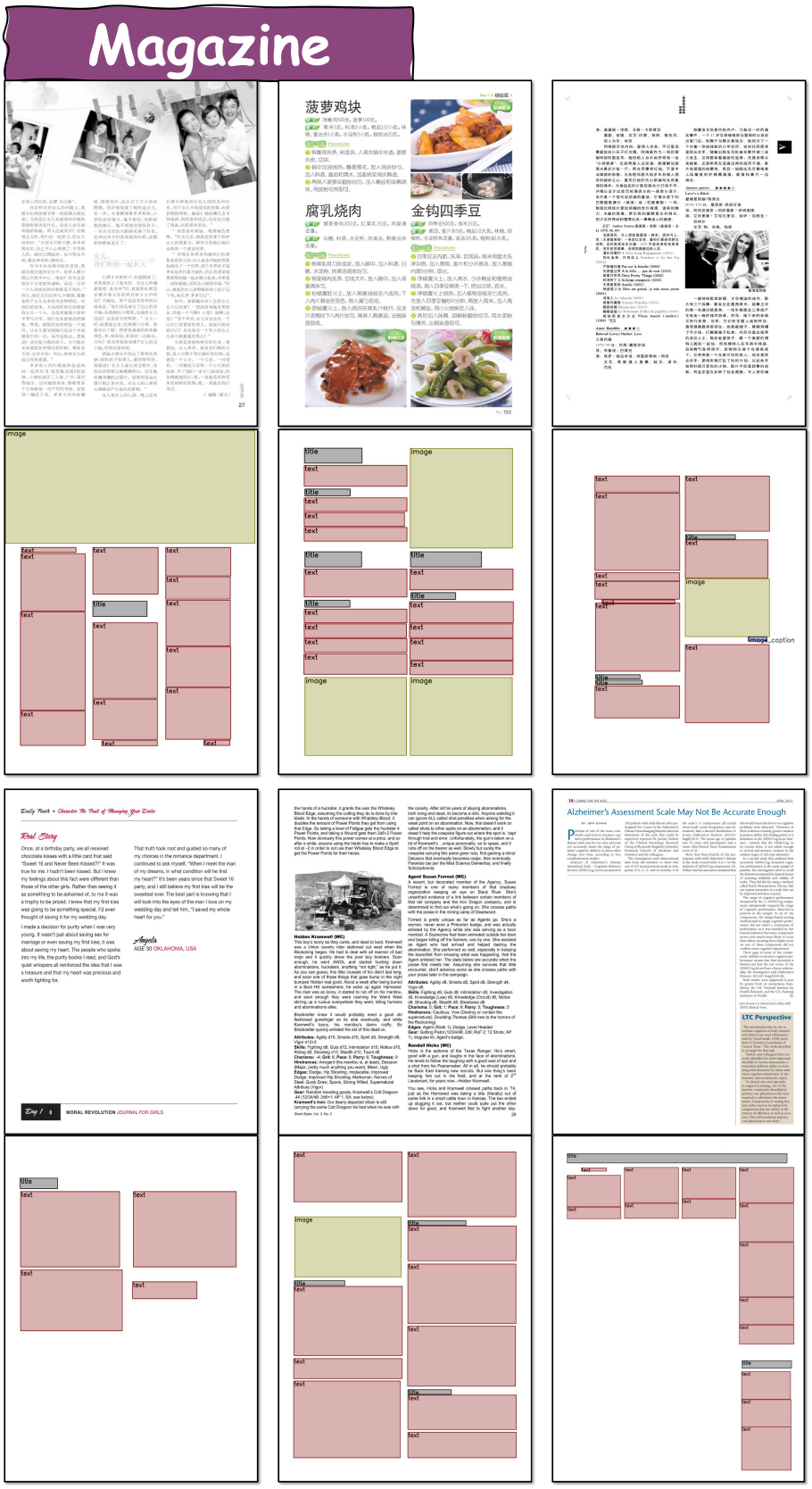


Figure 11: Visualization of Magazine Layout Data in OmniLayout-1M.

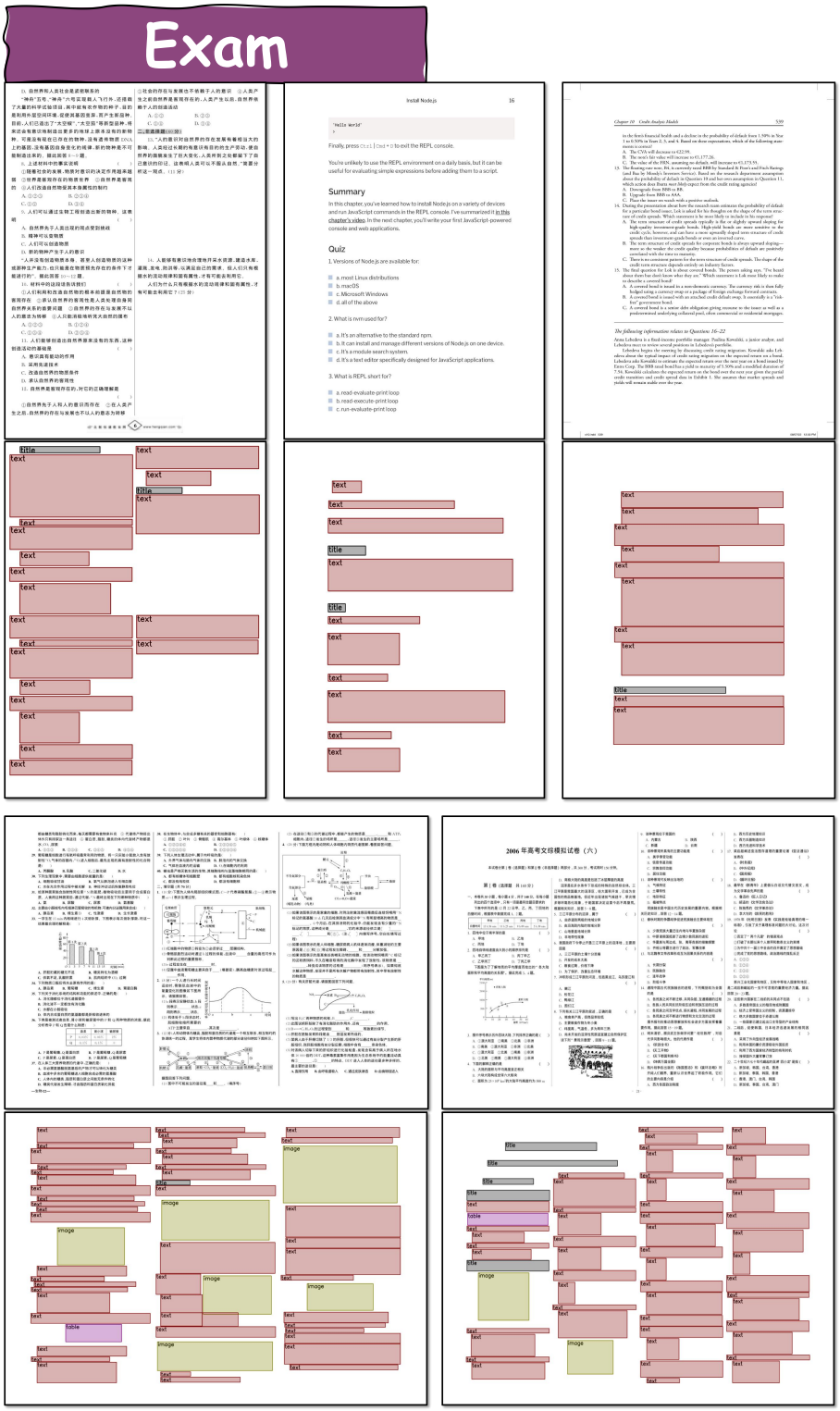


Figure 12: Visualization of Exam Layout Data in OmniLayout-1M.

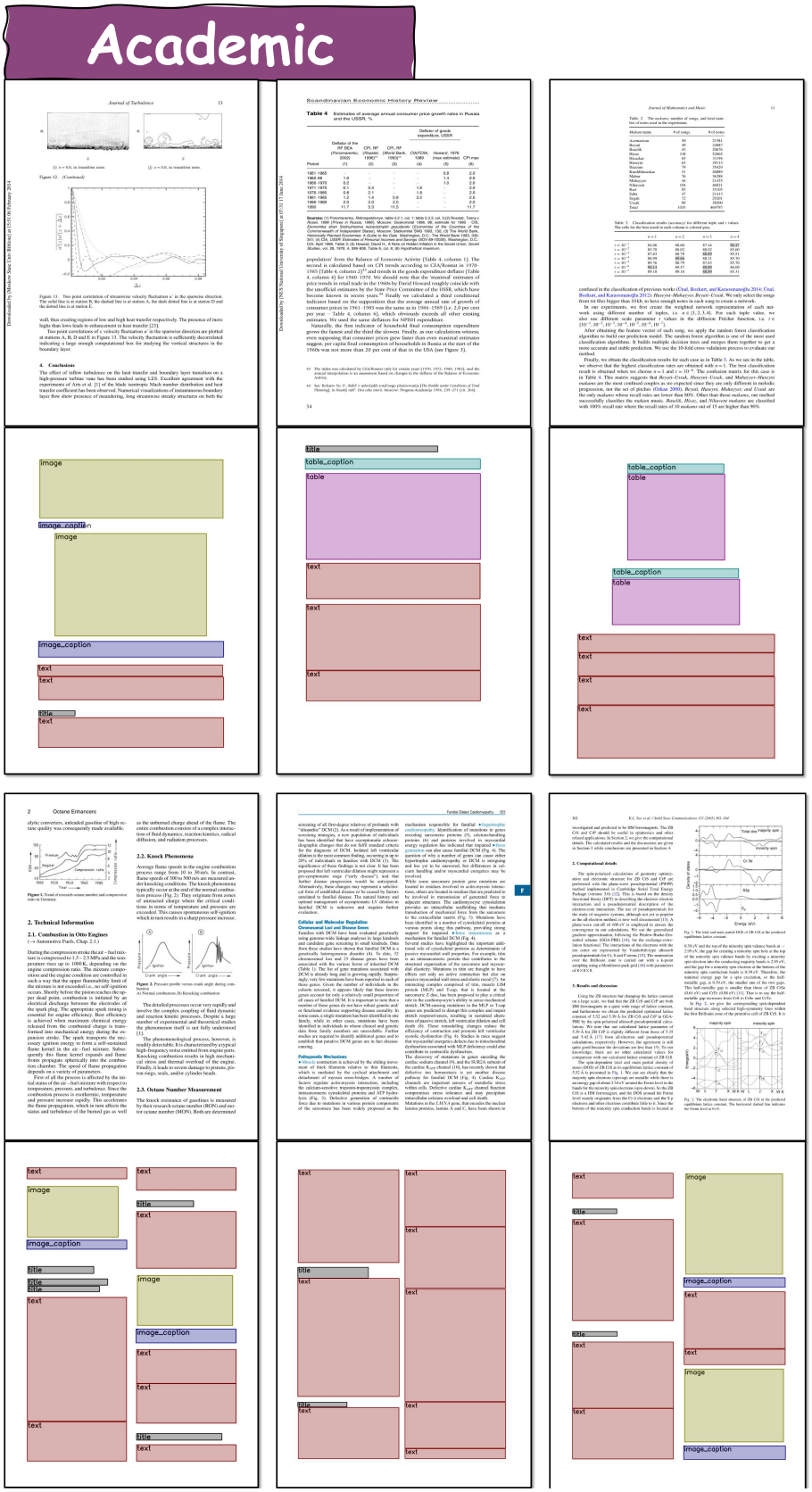


Figure 13: Visualization of Academic Layout Data in OmniLayout-1M.

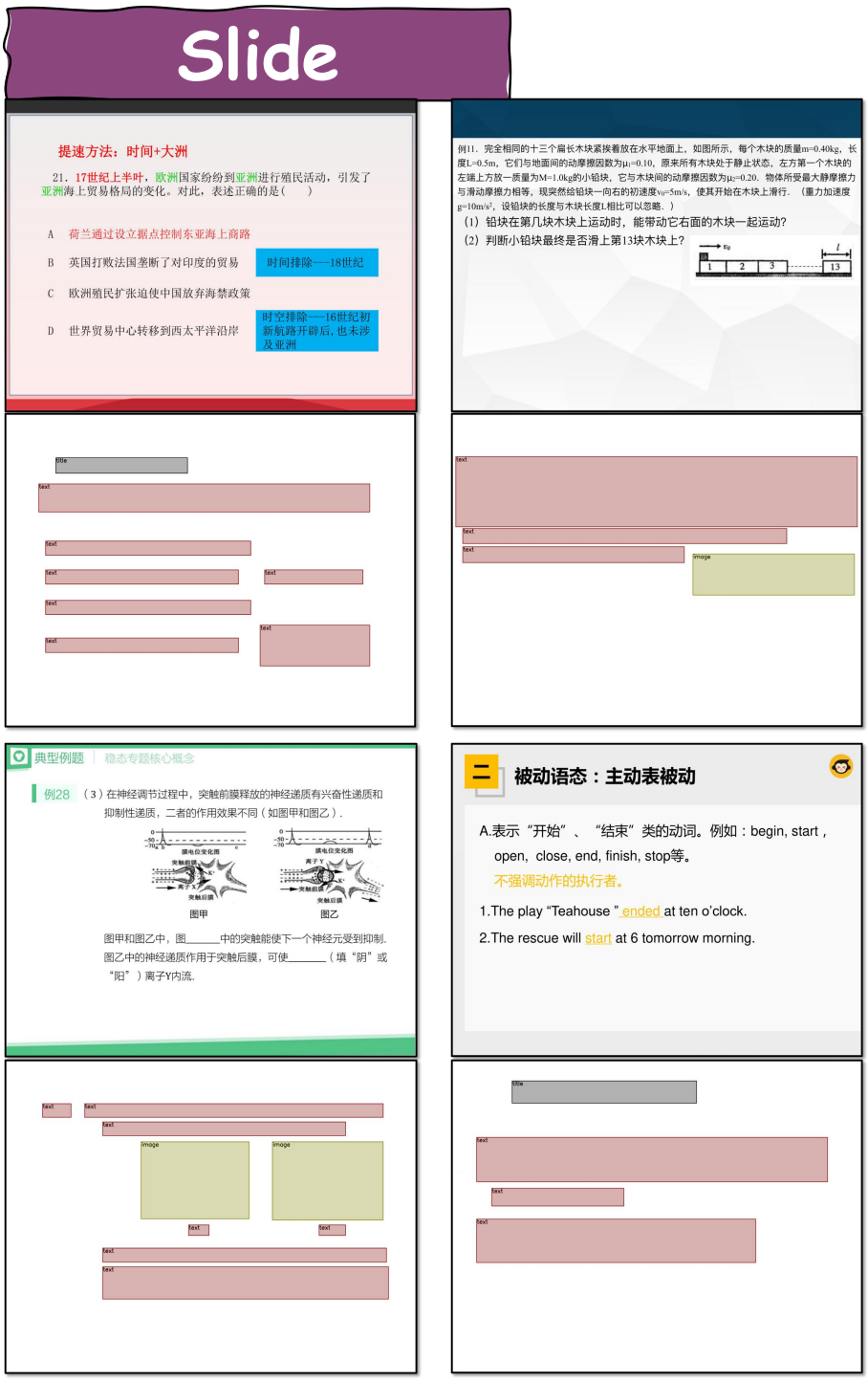
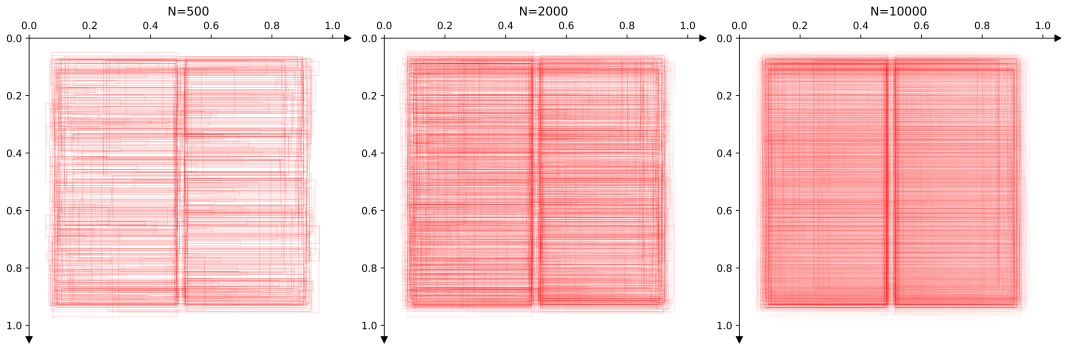
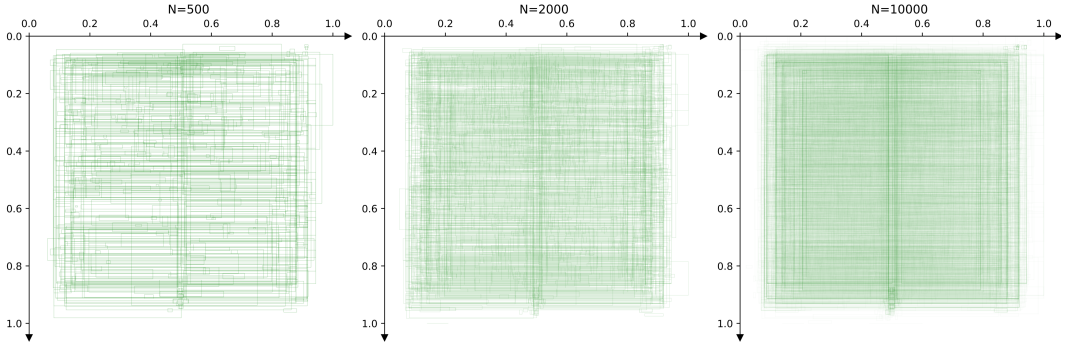


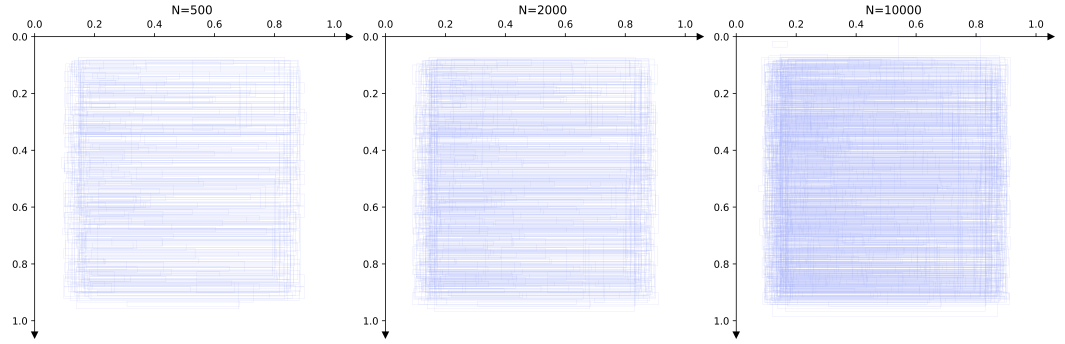
Figure 14: **Visualization of Slide Layout Data in OmniLayout-1M.**



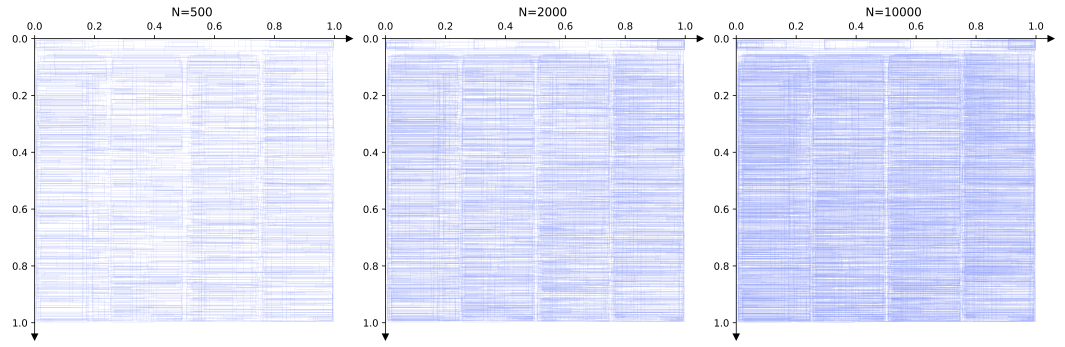
(a) Document layout distribution of PubLayNet.



(b) Document layout distribution of DocBank.

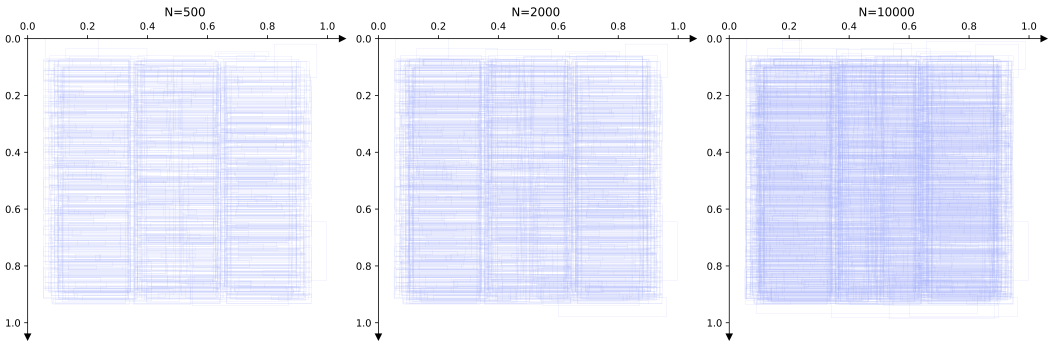


(c) Document layout distribution of textbook in OmniLayout-1M.

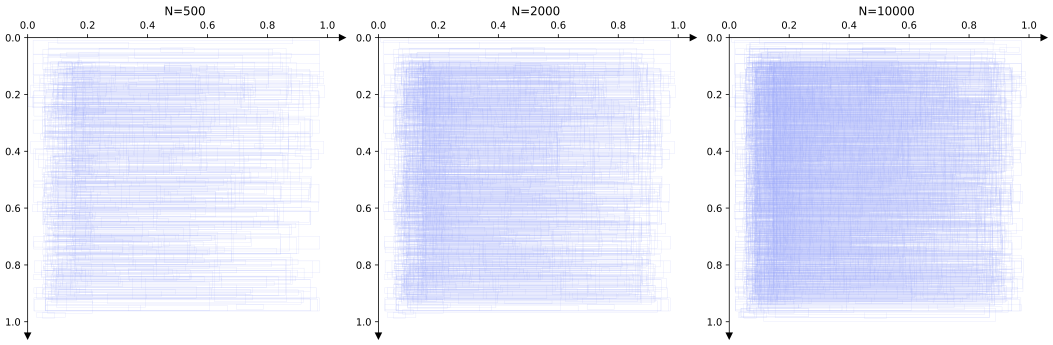


(d) **Document Layout distribution of newspaper in OmniLayout-1M.**

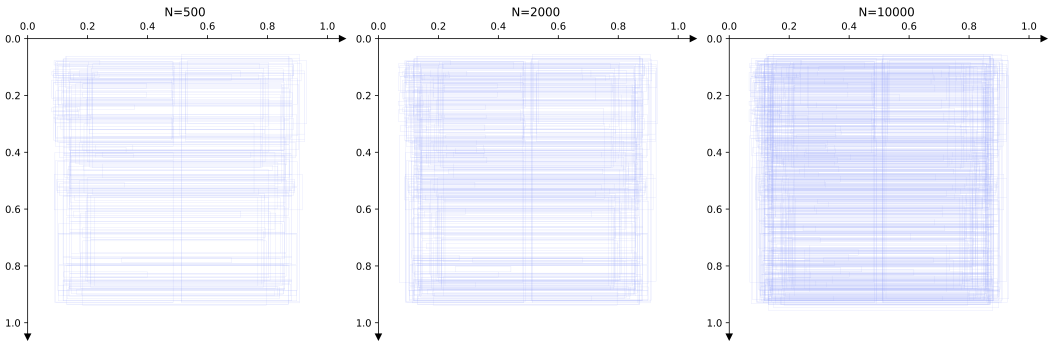
Figure 15: Document Layout Distribution of (a) PubLayNet, (b) DocBank, (c) Textbook in OmniLayout-1M, and (d) Newspaper in OmniLayout-1M.



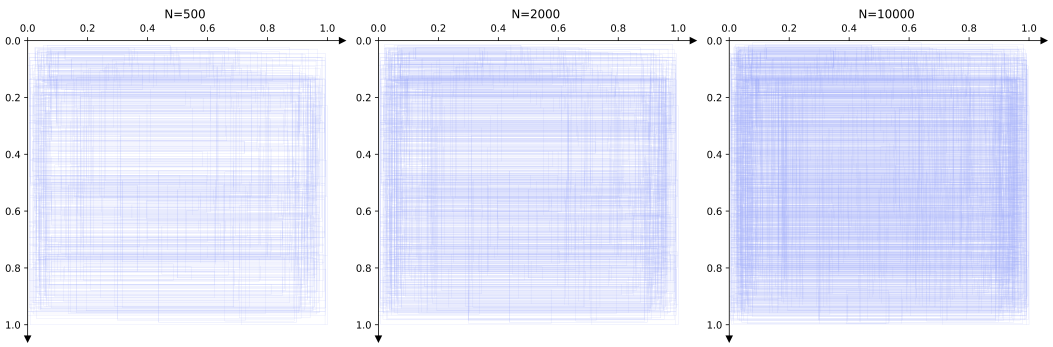
(a) Document layout distribution of magazine in OmniLayout-1M.



(b) Document layout distribution of exam in OmniLayout-1M.



(c) Document layout distribution of academic in OmniLayout-1M.



(d) Document layout distribution of slide in OmniLayout-1M.

Figure 16: Document Layout Distribution of (a) Magazine, (b) Exam, (c) Academic, and (d) Slide in OmniLayout-1M.