OmniDocLayout: Towards Diverse Document Layout Generation via Coarse-to-Fine LLM Learning

Supplementary Material

Contents of the Appendices:

A. Comparison of Scenario Scope and Layout Complexity with Similar Works

A.1. Comparison Across Multiple Scenario Scopes

Recently, layout generation has been extensively studied across diverse application scenarios. Existing approaches can be broadly categorized into three groups: room layout planning, graphic layout design, and document layout generation. To be specific, room layout planning [17, 31] focuses on partitioning indoor spaces into functional regions while satisfying spatial and accessibility constraints (*e.g.*, ensuring navigable and appropriately adjacent areas), typically involving a small and fixed set of element types. Graphic layout design (*e.g.*, poster [16, 60]), in contrast, emphasizes saliency-aware placement of a few elements such that the main subject of a background image remains unobstructed. Our work instead targets document layout generation, where numerous heterogeneous elements (*e.g.*, headers, paragraphs, tables, figures) must be arranged within a rectangular page according to the stylistic norms of different document types. Among these three application scenarios, as illustrated in Fig. 5, document layout generation is the most challenging, featuring:

- richer element categories, deeper hierarchical structures, and substantial style diversity across document types;
- stricter layout rules across multiple dimensions, including alignment, reading order, spacing, and style;
- significant challenges in modeling and spatial organization when handling a large number of elements per page.

Figure 5. Scenario Scope Comparison. (Left) shows a generated room layout by [31]. (Middle) shows a generated graphic layout by [16]. (Right) shows a generated document layout by our OmniDocLayout-LLM.

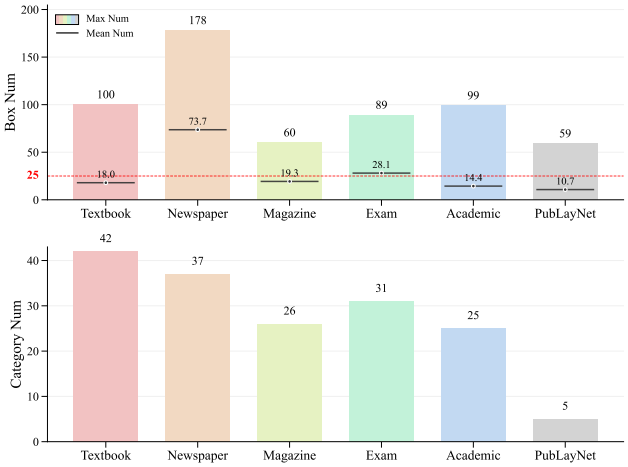


Figure 6. Layout Complexity Comparison Between Five Complex Types in M^{6}Doc and Widely-used PubLayNet. (Top) compares the maximum and average number of elements per page. (Bottom) compares the granularity of element categorization. The red dashed line indicates the default maximum number of elements (25) allowed in prior methods.

A.2. Comparison of Layout Complexity Across Various Documents

In this section, we examine prior work on document layout generation. Existing studies [4, 11, 19, 20, 26, 37, 47] primarily evaluate models on PubLayNet [59], which largely consists of simple layouts from single- or double-column academic papers. Our goal is to expand layout generation to a broader range of highly complex documents, and we therefore use the M^{6}Doc [5] dataset for evaluation. We compare our test set with PubLayNet from two perspectives, as illustrated in Figure 6. (1) Maximum and average number of elements per page. The five document types in M^{6}Doc consistently exhibit higher complexity in both metrics. In particular, the newspaper type reaches a maximum of 178 elements and an average of 73.7 elements per page, while PubLayNet contains only 59 elements at maximum and 10.7 on average. (2) Granularity of element categories. PubLayNet defines only 5 element types: text, title, list, table, and figure. In contrast, M^{6}Doc provides much finer-grained semantic categories. The textbook category contains 42 element classes, and even the academic category includes 25 classes. These findings indicate that our evaluation setting involves substantially higher layout complexity and generation difficulty than previous benchmarks, while also better reflecting real-world document diversity.

Notably, several recent studies, such as LayoutNUWA [38] and LGGPT [56], have also attempted to ex-

tend layout generation to more diverse and complex document styles. However, the range of document types they cover remains substantially narrower than ours, indicating that their evaluation settings are still less comprehensive.

B. OmniDocLayout-1M Dataset

B.1. Curation Details

Dual-deduplication. To remove duplicate samples from the final dataset, we adopted a *dual-deduplication strategy*: image-level deduplication was first applied during preprocessing, followed by layout-level deduplication after annotation. For image-level deduplication, we employed perceptual hashing (pHash) for document image deduplication. Each image was converted to grayscale and resized to 32x32 pixels, generating a 1024-bit hash. We compared hash values using Hamming distance, with a normalized threshold of 0.05, allowing for a certain tolerance of noise while identifying duplicates or near-duplicates. This resolution was chosen because document images contain structured layouts and fine textual details, and a higher number of bits provides a better representation to achieve more reliable accuracy. For layout-level deduplication, we calculated the mIoU (max Intersection over Union) score between each pair of samples. For pairs with an mIoU greater than 0.95, we retained only one sample selected at random.

Data Filtering. To ensure the quality of annotations from the automated labeling pipeline and discard low-quality samples that do not conform to prior knowledge, we performed a filtering step after annotation. We filtered the samples based on three dimensions: (1) Minimum number of elements. Based on empirical intuition, for documents such as textbooks, magazines, exams, and academic papers, we discarded pages with fewer than 5 elements, while for more complex documents like newspapers, this threshold was increased to 10. (2) Fill ratio. For certain densely arranged document types, such as newspapers, large amounts of white space are generally not allowed in the layout. Therefore, we filtered out pages with a fill ratio of less than 60%. (3) Overlap Score. Although overlaps are common in UI or poster design tasks, they are generally avoided in document layouts. Therefore, samples with abnormally high overlap scores were marked as incorrectly annotated and discarded. As for alignment score, it can be naturally high in certain non-Manhattan layout document types and is thus not suitable as a filtering criterion, as using it would lead to a significant loss of layout diversity.

B.2. Final Quality Control

To address the concern that using model-generated bounding boxes as training data may introduce annotation noise or bias, we design a blind human evaluation to directly assess the perceptual quality of model-annotated layouts relative to human-annotated ground truth.

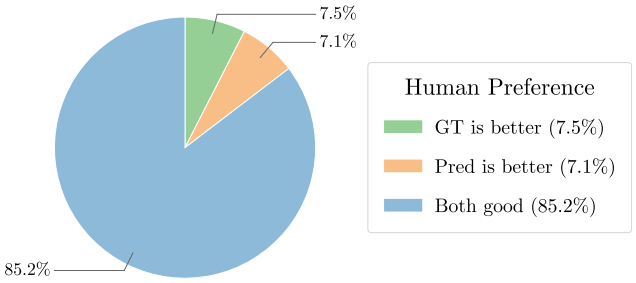


Figure 7. Human evaluation of 1,200 pages comparing model-generated and human-annotated layouts. "Pred" refer to the bounding box predicted by MinerU [39].

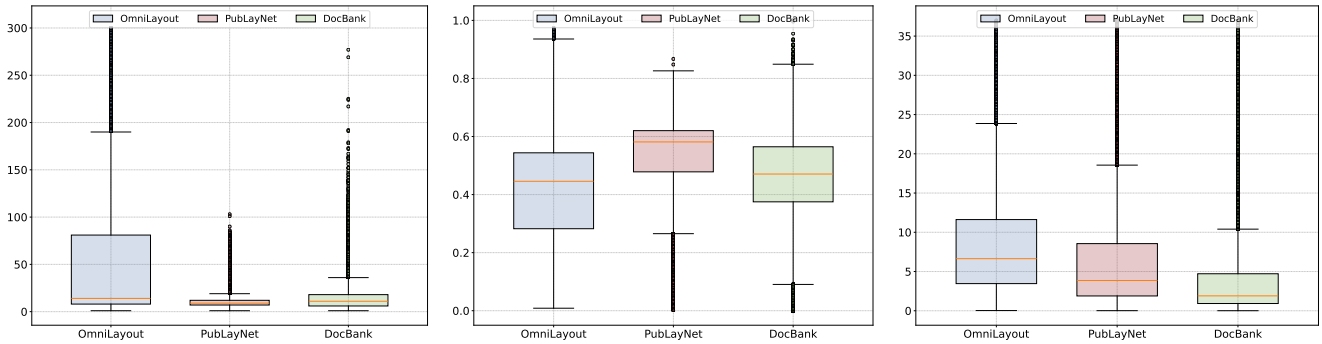
Setup. We randomly sample 1,200 document pages from the OmniDocLayout-1M dataset, selecting 200 pages for each document type (textbook, newspaper, magazine, exam, academic and slide), and obtain two sets of annotations for each page: (*i*) human-annotated bounding boxes (GT), manually labeled by trained annotators following our specification, and (*ii*) model-generated bounding boxes produced by our layout extraction model without post-processing. For every page, we construct a pair of annotations (A, B), where one corresponds to the human annotation and the other to the model output, ensuring fully unbiased comparisons, with their order randomized and hidden from evaluators.

Evaluation Procedure. The evaluation interface displays the document page image alongside two independently rendered layout annotations (A and B). Annotators are asked to decide whether A is better, B is better, or both are good. Selecting "Both good" indicates that the two annotations are both accurate and visually aligned with the page structure. All evaluators are familiar with document layout quality criteria but are not informed which annotation is the ground truth or the model prediction produced by MinerU [39]. In total, four annotators participated in the study, and each sample is reviewed by two annotators, with disagreements resolved through majority voting.

Result Analysis. Across all 1,200 evaluated pages, the distribution of votes in Fig. 7 shows that 1,022 samples (85.2%) fall into the "Both good" category, indicating that human and model-generated layouts are largely indistinguishable in perceptual quality. An additional 85 samples (7.1%) are judged as "Model is better," and together these outcomes account for over 92% of cases where model-generated bounding boxes are considered at least as good as human annotations. These results demonstrate a high level of human–model consistency in layout quality.

B.3. Element-level Statistical Analysis

First, we analyze the diversity of OmniDocLayout-1M from the perspective of element distribution. Specifically, we examine element diversity in three aspects: the number of el-



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

Figure 8. Element-Level Statistical Comparison Between Our OmniDocLayout-1M and Two Widely-Used Datasets: PubLayNet and DocBank.

ements per page, the proportion of the layout area occupied by all elements on a single page, and the aspect ratios of the elements. The data distribution is illustrated in Fig. 8. As can be observed, OmniDocLayout-1M exhibits significantly greater diversity in elements compared to PubLayNet [59] and DocBank [24]. 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.

B.4. Layout Diversity

Next, we visualize and compare the document layout diversity of PubLayNet, DocBank, and OmniDocLayout-1M as shown in Fig. 9 and Fig. 10. N indicates the 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 OmniDocLayout-1M significant variation and diversity.

B.5. More Visual Examples

In this section, we present more visual examples from our OmniDocLayout-1M dataset, accompanied by high-quality annotations extracted with MinerU [39]. We visualize annotated examples of six document types: textbook (Fig. 11), newspaper (Fig. 12), magazine (Fig. 13), exam (Fig. 14), academic (Fig. 15), slide (Fig. 16) are shown.

C. OmniDocLayout-LLM Analysis

C.1. Few-shot Performance of General-purpose LLMs

Due to the page limit, 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.

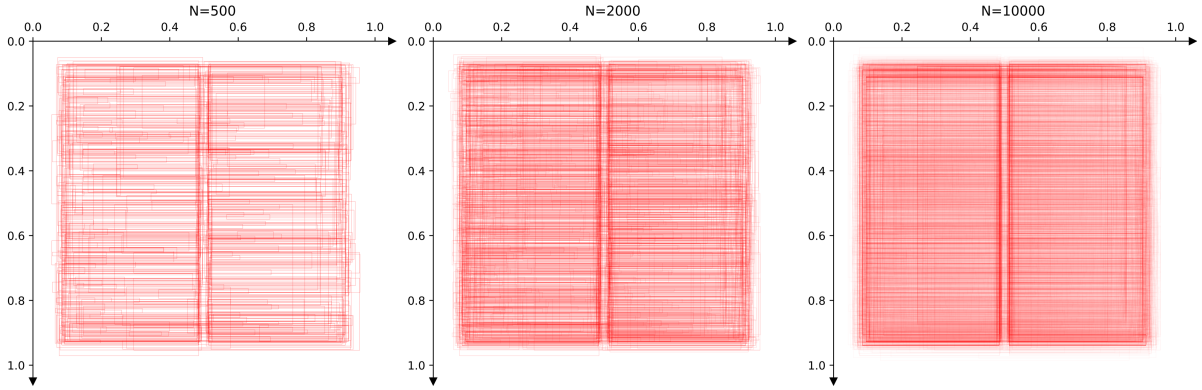
General-purpose LLMs will overtake? Experimental results have shown improved performance with increasing numbers of in-context examples, raising concerns about whether future versions of such models, or simply scaling the number of shots, could replace specialized models in this scenario. It is a challenge faced by nearly all task-specific research areas. However, fully relying on general-purpose LLMs or in-context learning to take over document layout generation remains difficult for two key reasons: (1) Lack of task-specific pretraining and limited spatial reasoning. Our zero-shot experiments show that general models often produce layouts with extensive overlaps, even when explicitly instructed to avoid them. This indicates that document layout generation is still a niche task, and current or future general models are unlikely to undergo targeted pretraining for this domain. Furthermore, these models exhibit limited spatial reasoning capabilities, which are essential for generating well-organized layouts. (2) Unacceptable inference-time and financial overhead when scaling shots. Increasing the number of in-context examples does not scale effectively in practice. This is because each complex layout will be serialized into a long sequence, and a 5-shot prompt already results in substantial token length, inference latency, and API cost. Scaling to 10 shots or more becomes impractical due to both inference time and financial cost. Together, these observations indicate that, while general-purpose LLMs show promising flexibility, they are unlikely to replace domain-specific approaches for document layout generation in the foreseeable future.

C.2. More Generated Layouts

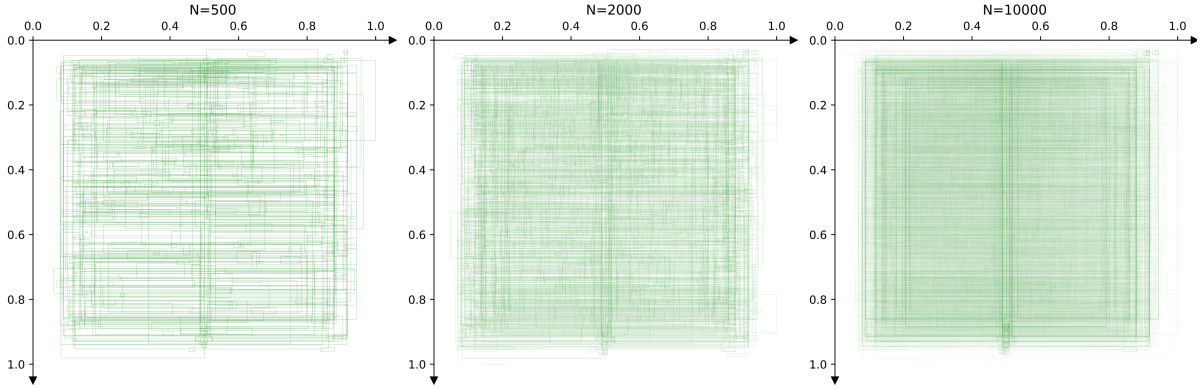
In this section, we demonstrate more qualitative examples generated by OmniDocLayout-LLM in five generation tasks: U-Cond, C→S+P, C+S→P, Complement and Refinement. Using M^{6}Doc as the test set, we show the generated layouts in Fig. 17 for textbooks and newspapers, Fig. 18 for magazines and exams, and Fig. 19 for academic papers. The visualization results indicate that our model can generate both reasonable and aesthetically pleasing layouts for diverse and complicated 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.

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

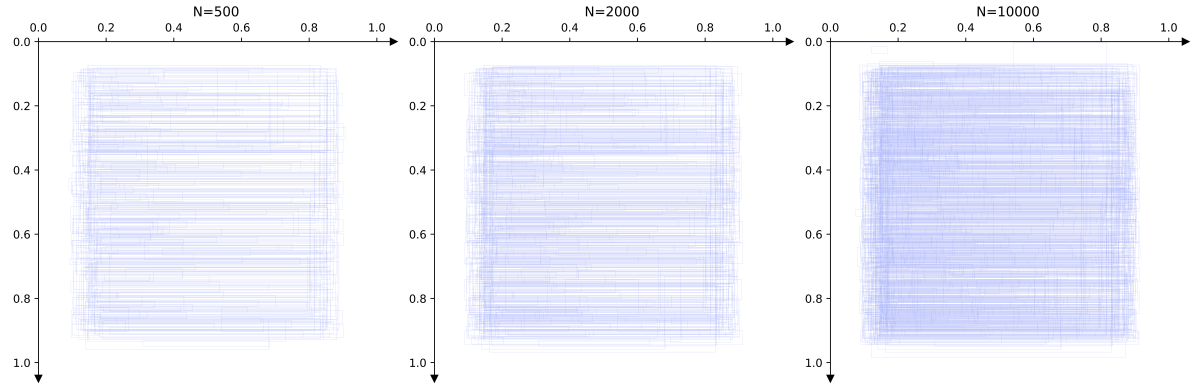
Task	Method	Setting	Textbook				Newspaper				Magazine				Exam				Academic			
			FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑	FID↓	Ali.→	Ove.→	mIoU↑
U-Cond	GPT-4o	0-shot	135.32	0.017	**0.007**	0.060	193.13	0.020	0.007	0.000	236.11	0.015	**0.040**	0.089	163.94	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	0.169	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	0.166	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	



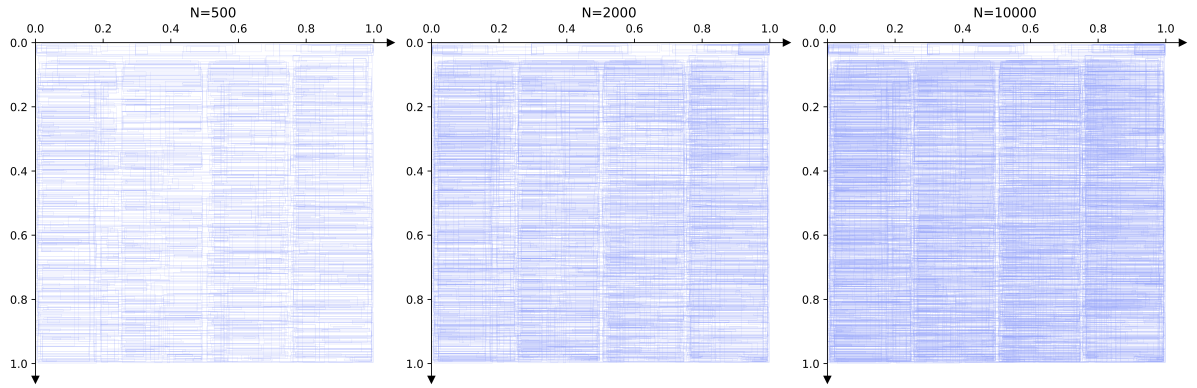
(a) Document layout distribution of PubLayNet.



(b) Document layout distribution of DocBank.

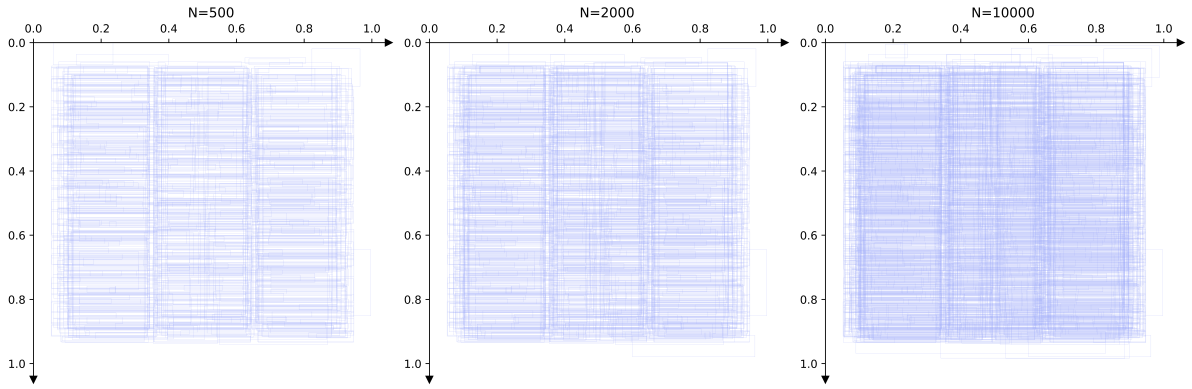


(c) Document layout distribution of textbook in OmniDocLayout-1M.

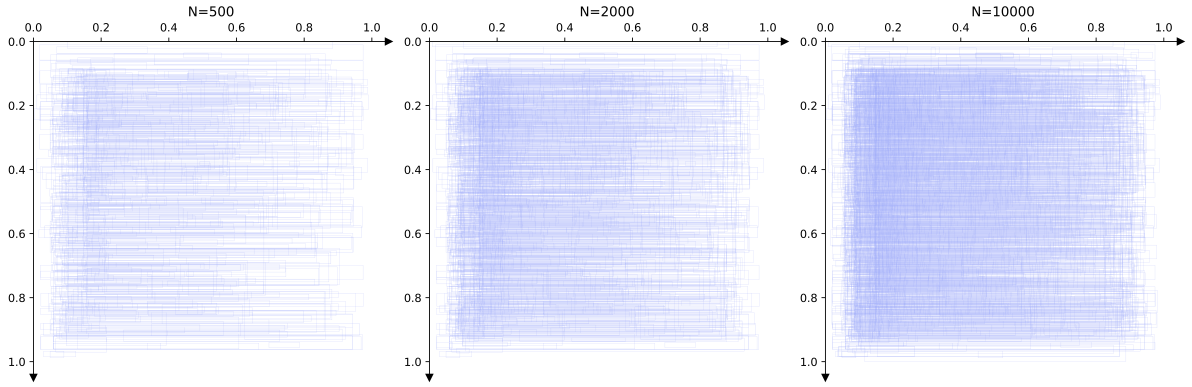


(d) Document layout distribution of newspaper in OmniDocLayout-1M.

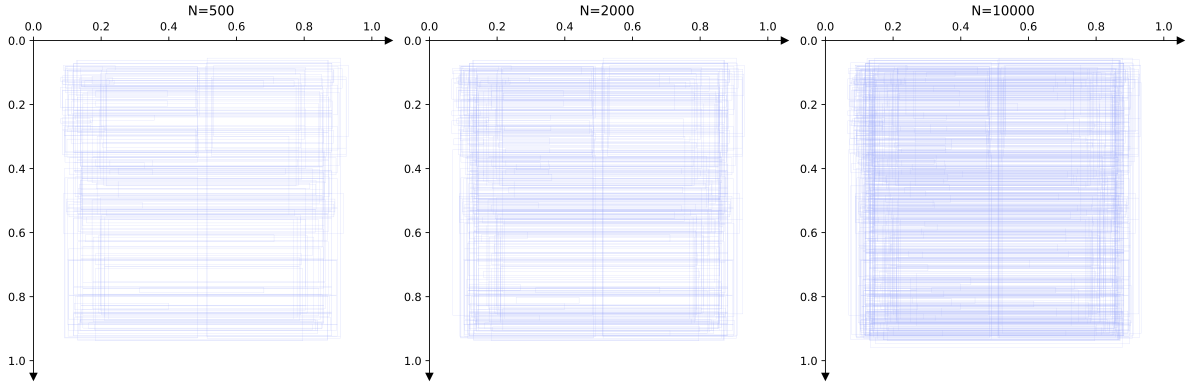
Figure 9. Document Layout Distributions of Two Widely-used Benchmarks, (a) PubLayNet and (b) DocBank, and Two Document Types from Our OmniDocLayout-1M Dataset, (c) Textbook and (d) Newspaper.



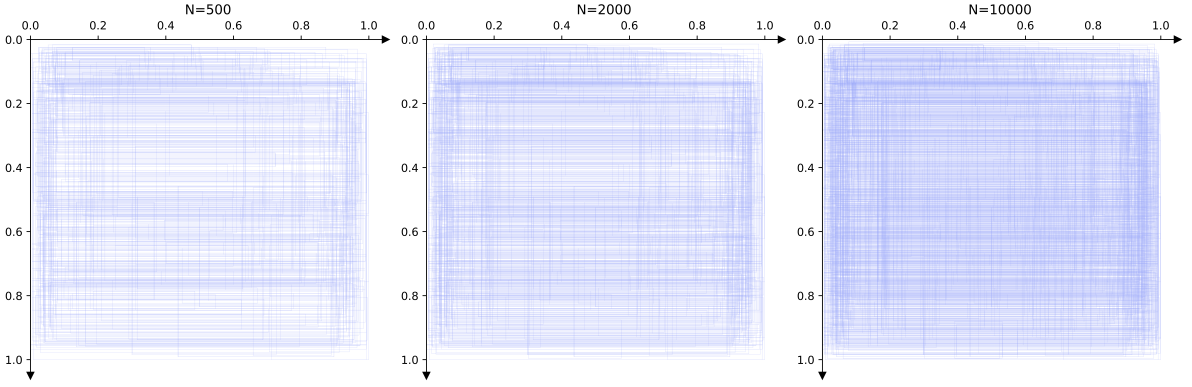
(a) Document layout distribution of magazine in OmniDocLayout-1M.



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



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



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

Figure 10. Document Layout Distributions of (a) Magazine, (b) Exam, (c) Academic, and (d) Slide in Our OmniDocLayout-1M Dataset.

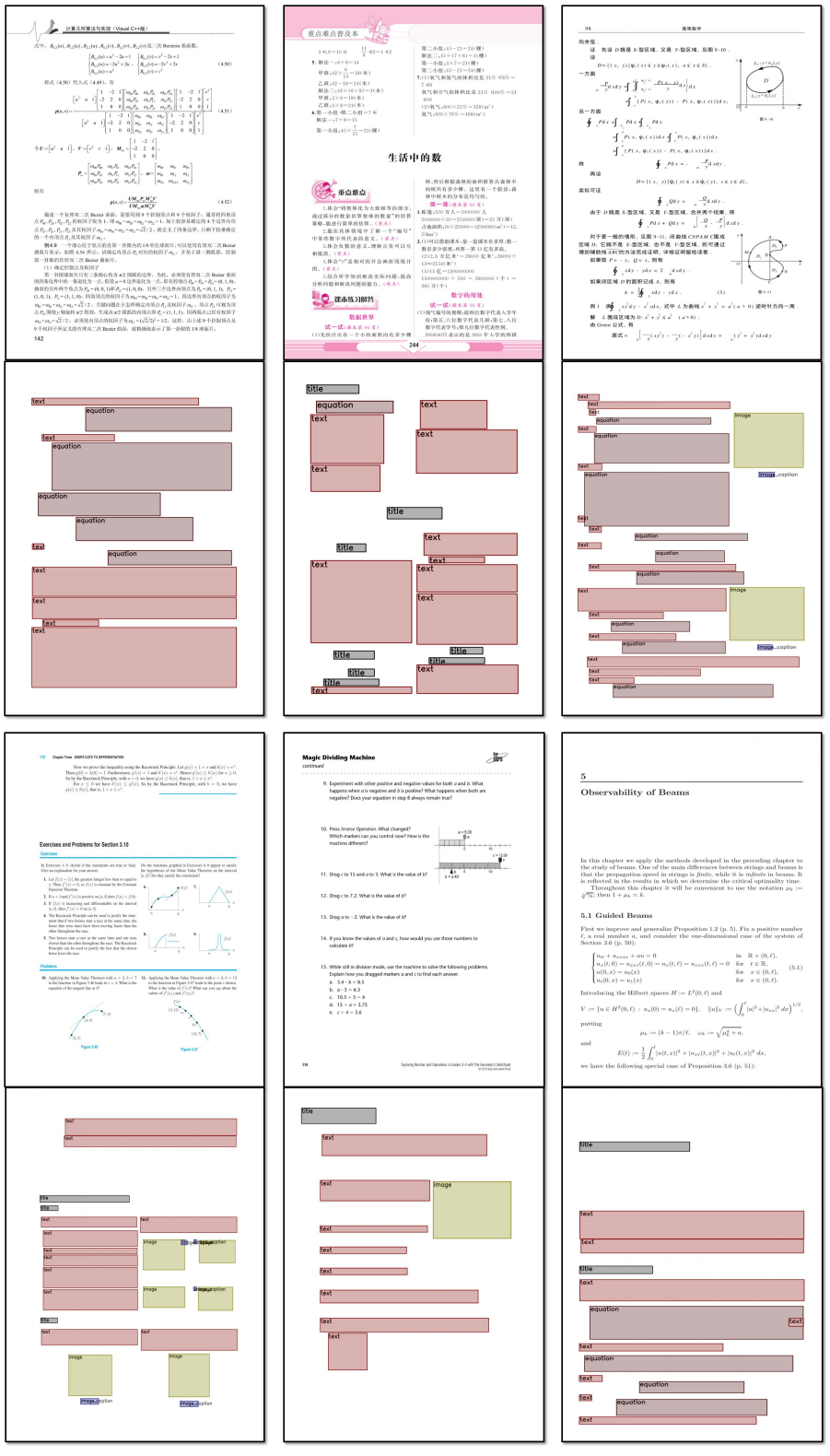


Figure 11. Example Textbook Layouts from Our OmniDocLayout-1M Dataset.

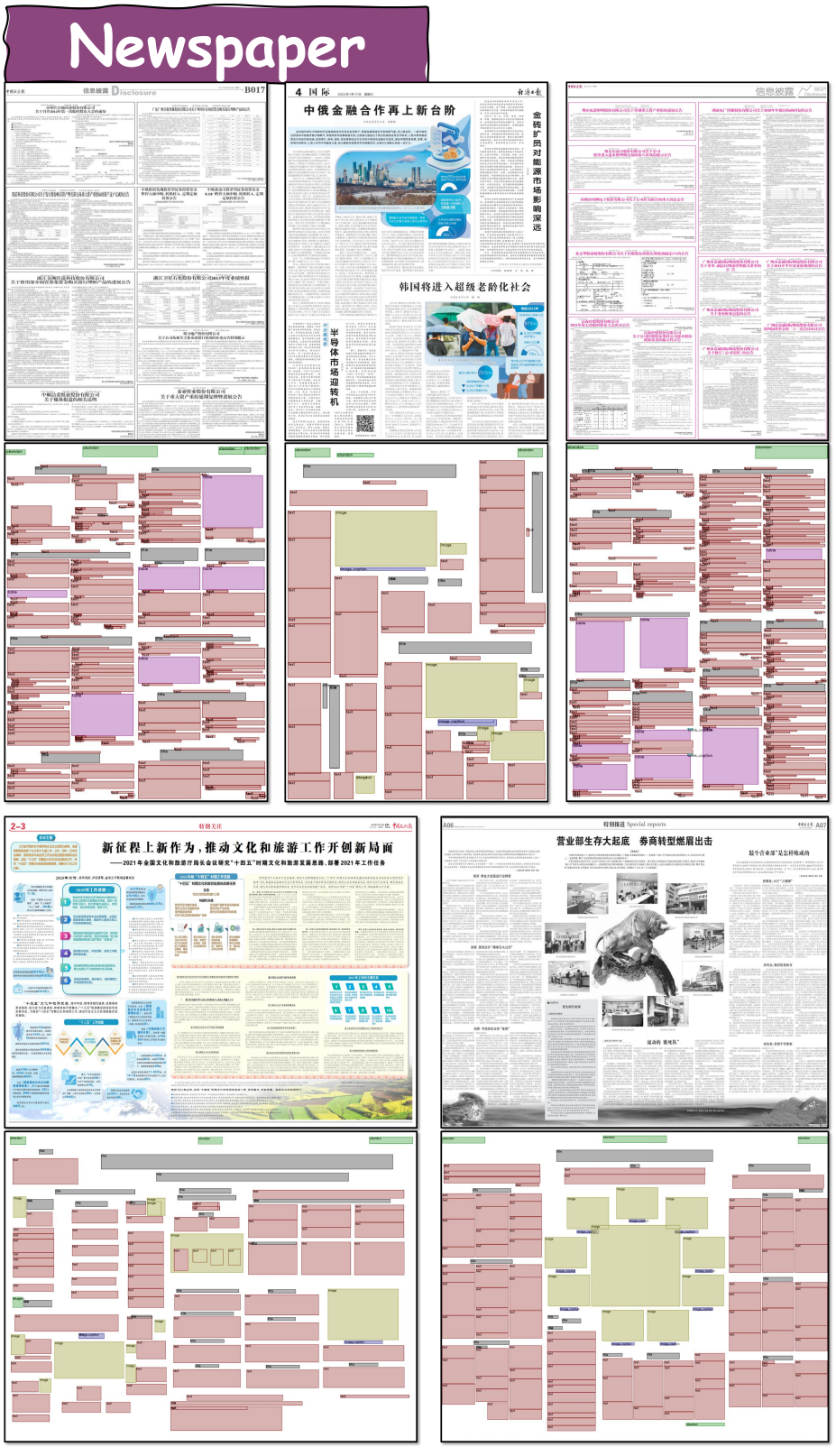


Figure 12. Example Newspaper Layouts from Our OmniDocLayout-1M Dataset.

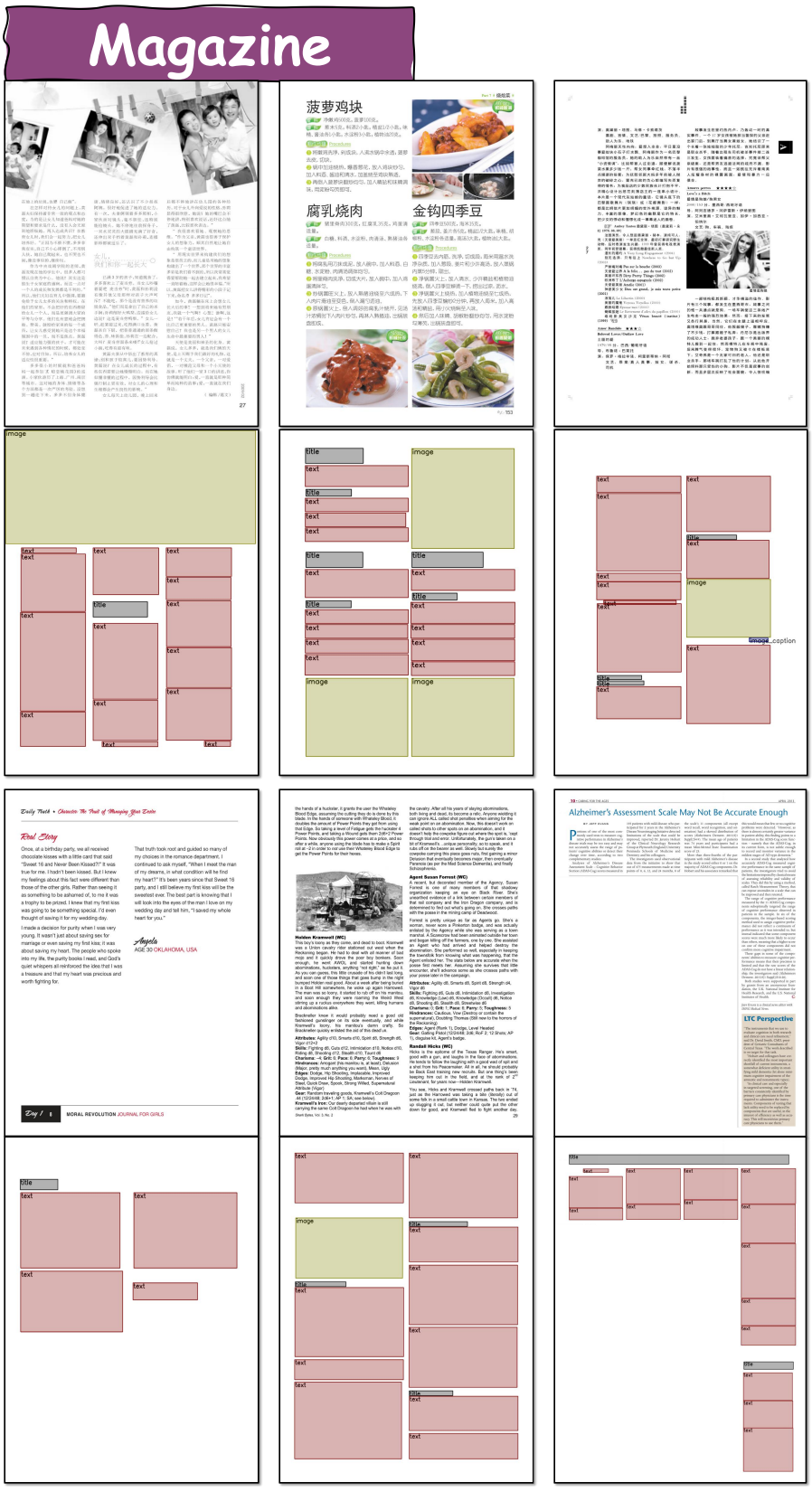


Figure 13. Example Magazine Layouts from Our OmniDocLayout-1M Dataset.

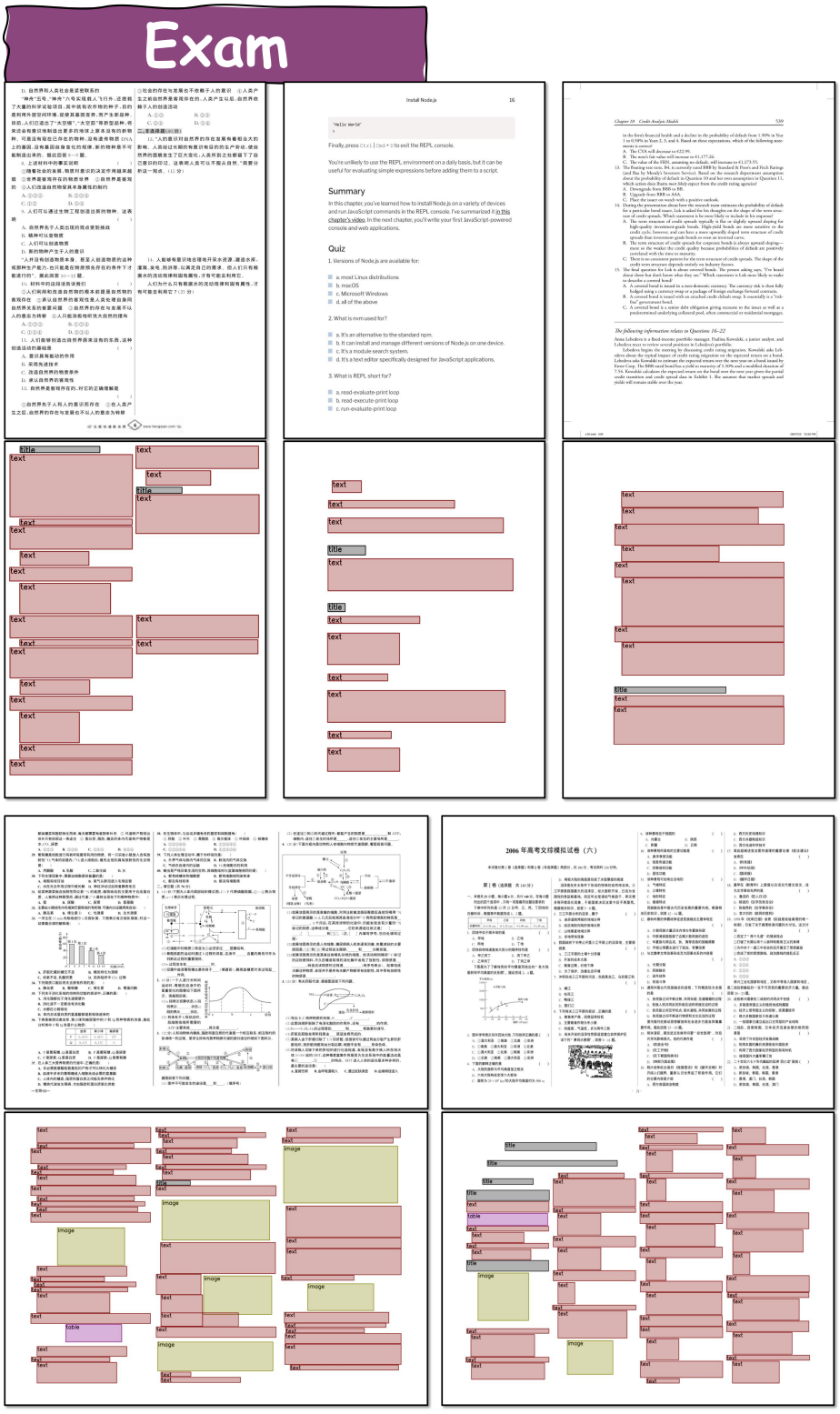


Figure 14. Example Exam Layouts from Our OmniDocLayout-1M Dataset.

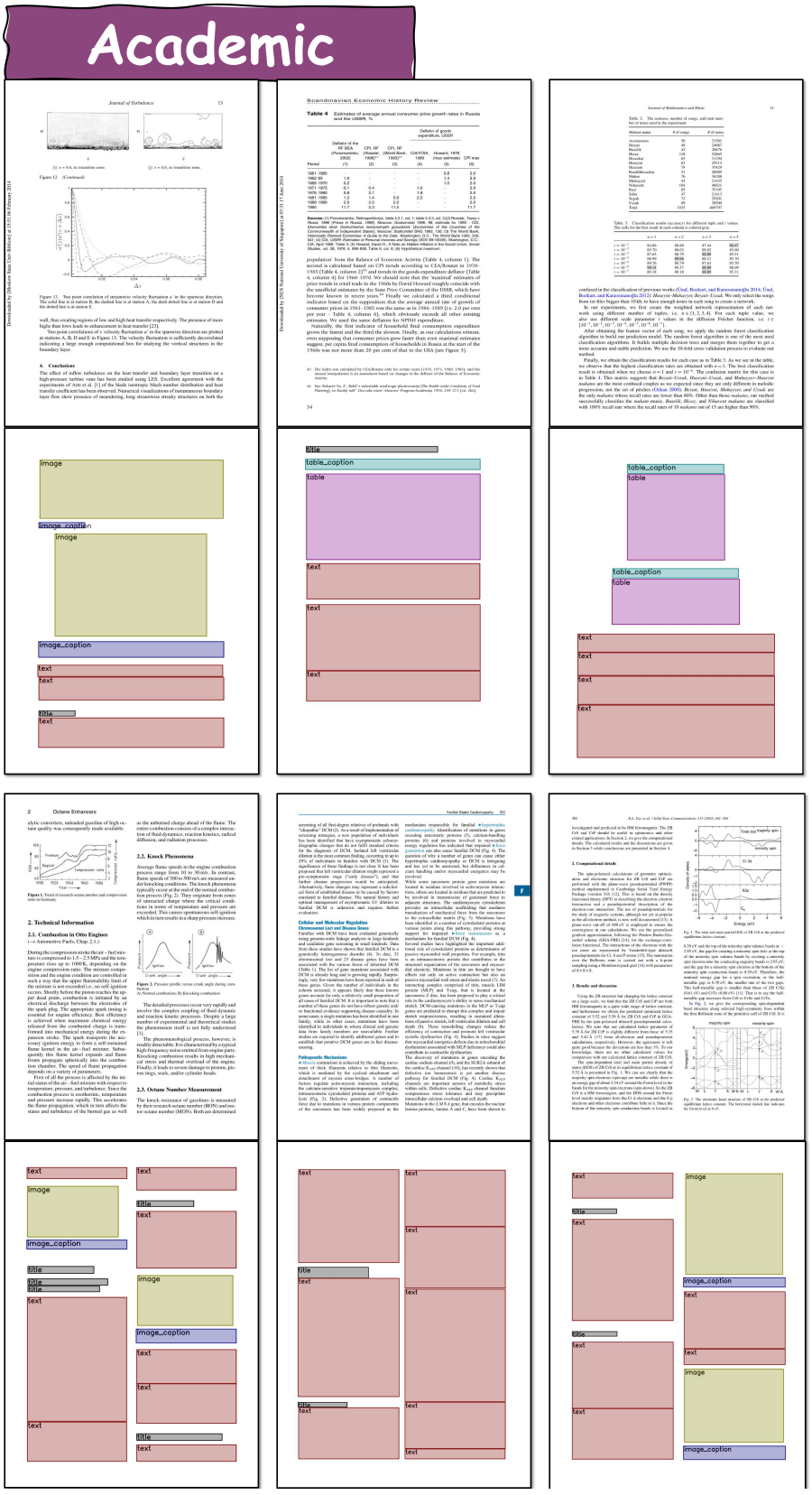


Figure 15. Example Academic Layouts from Our OmniDocLayout-1M Dataset.

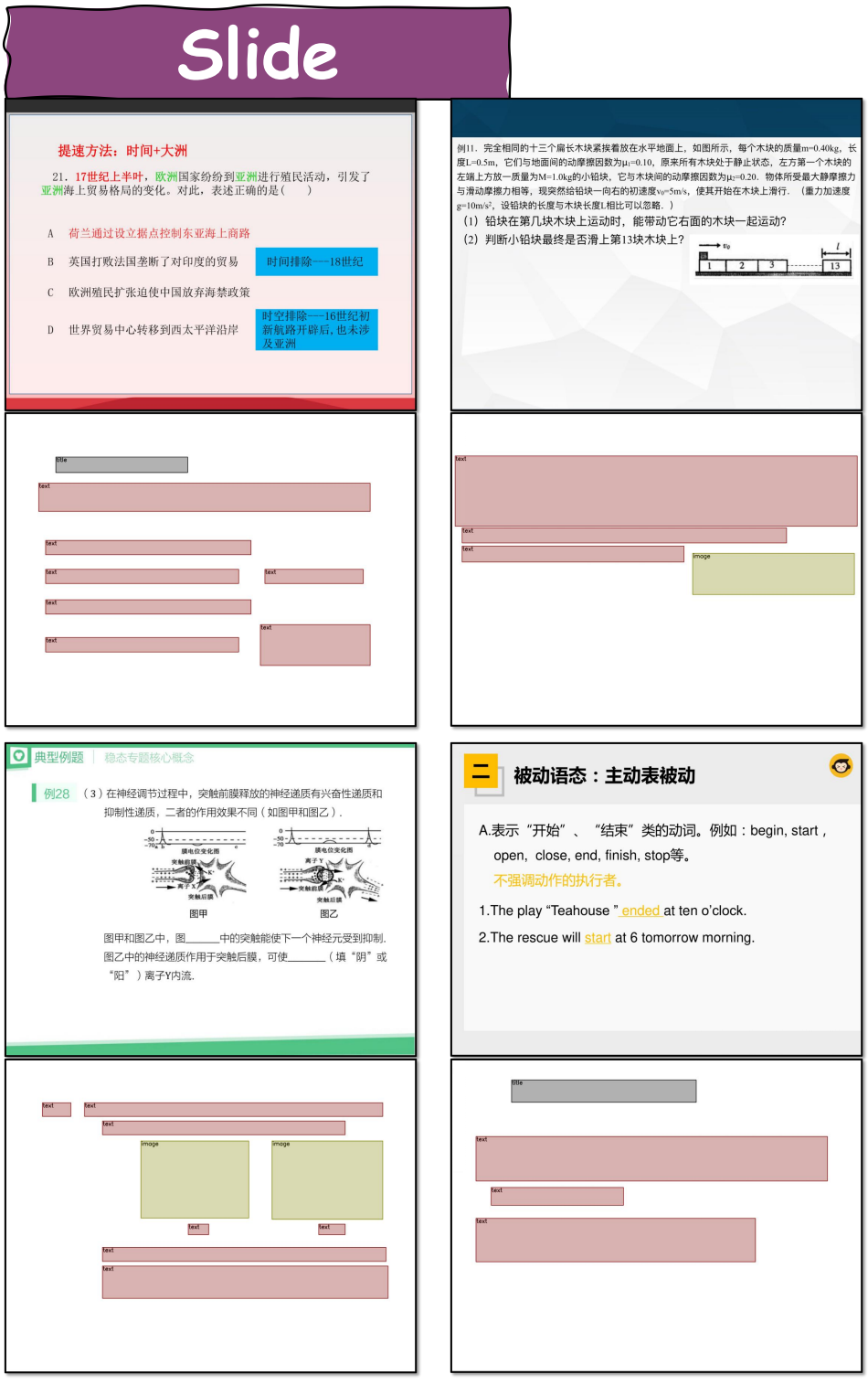


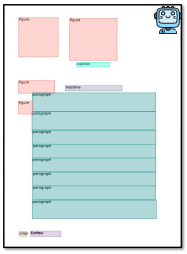
Figure 16. Example Slide Layouts from Our OmniDocLayout-1M Dataset.

U-COND

Base Prompt
Document Type: "textbook"
Canvas Size: [689, 1000]
Bbox Number: 16
Valid Categories: {"answer", "author", "blank", ...}

Condition Prompt
None

Task Prompt
Please perform the unconditional layout generation task based on the above information. Specifically, the unconditional task does not provide any existing bbox information, and you need to generate the entire layout from scratch....

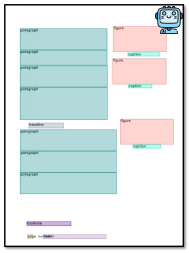


C→S+P

Base Prompt
Document Type: "textbook"
Canvas Size: [707, 1000]
Bbox Number: 17
Valid Categories: {"answer", "author", "blank", ...}

Condition Prompt
<|cat_start|>**paragraph**<|cat_end|><|box_start|>**figure**<|cat_end|><|box_start|>;<|cat_start|>**caption**<|cat_end|><|box_start|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>;<|cat_start|>**paragraph**<|cat_end|>;...

Task Prompt
Please perform the c layout generation task based on the above information. Specifically, the c task provides the category information for each bbox, and you need to predict both the position coordinates and size information for them....

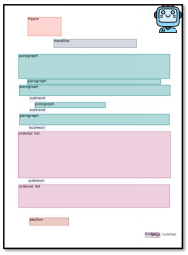


C+S→P

Base Prompt
Document Type: "textbook"
Canvas Size: [676, 1000]
Bbox Number: 16
Valid Categories: {"answer", "author", "blank", ...}

Condition Prompt
<|cat_start|>**page number**<|cat_end|><|box_start|>**2033 3020**<|box_end|>;<|cat_start|>**footer**<|cat_end|><|box_start|>**2051 3020**<|box_end|>;<|cat_start|>**figure**<|cat_end|><|box_start|>**2127...**

Task Prompt
Please perform the cwh layout generation task based on the above information. Specifically, the cwh task provides the category and size information for each bbox, and you need to predict the position coordinates for them....

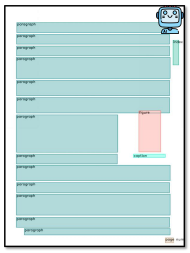


COMPLETION

Base Prompt
Document Type: "textbook"
Canvas Size: [684, 1000]
Bbox Number: 17
Valid Categories: {"answer", "author", "blank", ...}

Condition Prompt
<|cat_start|>**paragraph**<|cat_end|><|box_start|>**41 1663 2591 3065**<|box_end|>;<|cat_start|>**figure**<|cat_end|><|box_start|>**512 1436 2084 3172**<|box_end|>;<|cat_start|>**caption**<|cat_end|>

Task Prompt
Please perform the completion layout generation task based on the above information. Specifically, the completion task provides partial bboxes, and you need to complete the remaining layout elements accordingly....

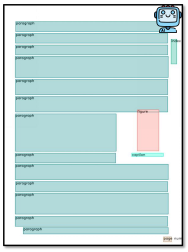


REFINEMENT

Base Prompt
Document Type: "textbook"
Canvas Size: [722, 1000]
Bbox Number: 17
Valid Categories: {"answer", "author", "blank", ...}

Condition Prompt
<|cat_start|>**unordered list**<|cat_end|><|box_start|>**109 1087 2269 3021**<|box_end|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>**85 1131 2532 3062**<|box_end|>;<|cat_start|>**unordered list**....

Task Prompt
Please perform the refinement layout generation task based on the above information. Specifically, the refinement task provides each bbox perturbed by noise, and you need to adjust and optimize them to improve the layout quality....



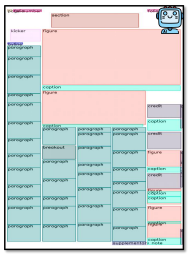
Textbook

U-COND

Base Prompt
Document Type: "newspaper"
Canvas Size: [499, 1000]
Bbox Number: 56
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
None

Task Prompt
Please perform the unconditional layout generation task based on the above information. Specifically, the unconditional task does not provide any existing bbox information, and you need to generate the entire layout from scratch....

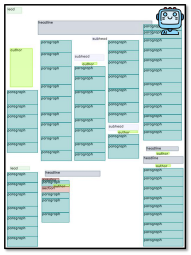


C→S+P

Base Prompt
Document Type: "newspaper"
Canvas Size: [707, 1000]
Bbox Number: 79
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
<|cat_start|>**headline**<|cat_end|><|box_start|>**subhead**<|cat_end|><|box_start|>**author**<|cat_end|><|box_start|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>;<|cat_start|>**paragraph**...

Task Prompt
Please perform the c layout generation task based on the above information. Specifically, the c task provides the category information for each bbox, and you need to predict both the position coordinates and size information for them....

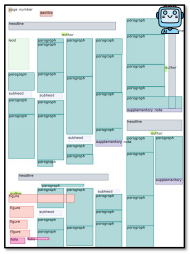


C+S→P

Base Prompt
Document Type: "newspaper"
Canvas Size: [707, 1000]
Bbox Number: 61
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
<|cat_start|>**headline**<|cat_end|><|box_start|>**2424 3033**<|box_end|>;<|cat_start|>**author**<|cat_end|><|box_start|>**2022 3007**<|box_end|>;<|cat_start|>**lead**<|cat_end|><|box_start|>**2082**...

Task Prompt
Please perform the cwh layout generation task based on the above information. Specifically, the cwh task provides the category and size information for each bbox, and you need to predict the position coordinates for them....

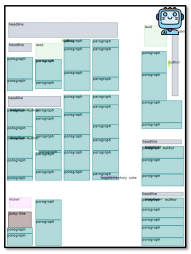


COMPLETION

Base Prompt
Document Type: "newspaper"
Canvas Size: [684, 1000]
Bbox Number: 53
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
<|cat_start|>**paragraph**<|cat_end|><|box_start|>**6 1211 2101 3090**<|box_end|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>**6 1541 2101 3090**<|box_end|>;<|cat_start|>**paragraph**<|cat_end|>...

Task Prompt
Please perform the completion layout generation task based on the above information. Specifically, the completion task provides partial bboxes, and you need to complete the remaining layout elements accordingly....

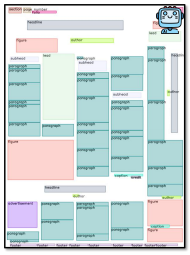


REFINEMENT

Base Prompt
Document Type: "newspaper"
Canvas Size: [722, 1000]
Bbox Number: 67
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
<|cat_start|>**unordered list**<|cat_end|><|box_start|>**109 1087 2269 3021**<|box_end|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>**85 1131 2532 3062**<|box_end|>;<|cat_start|>**unordered list**....

Task Prompt
Please perform the refinement layout generation task based on the above information. Specifically, the refinement task provides each bbox perturbed by noise, and you need to adjust and optimize them to improve the layout quality....



Newspaper

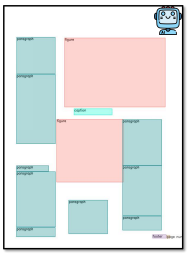
Figure 17. Examples of Layouts Generated by Our OmniDocLayout-LLM on Five Generation Tasks (Textbook and Newspaper).

U-COND

Base Prompt
Document Type: "magazine"
Canvas Size: [764, 1000]
Bbox Number: 14
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
None

Task Prompt
Please perform the unconditional layout generation task based on the above information. Specifically, the unconditional task does not provide any existing bbox information, and you need to generate the entire layout from scratch....

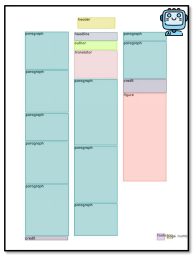


C→S+P

Base Prompt
Document Type: "magazine"
Canvas Size: [719, 1000]
Bbox Number: 19
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
<|cat_start|>**header**<|cat_end|><|box_start|>;<|cat_start|>**footer**<|cat_end|>;<|cat_start|>**pagenumber**<|cat_end|>;<|cat_start|>**paragraph**<|cat_end|>;...

Task Prompt
Please perform the c layout generation task based on the above information. Specifically, the c task provides the category information for each bbox, and you need to predict both the position coordinates and size information for them....

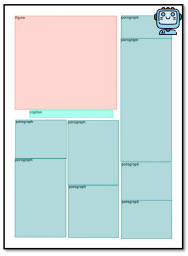


C+S→P

Base Prompt
Document Type: "magazine"
Canvas Size: [733, 1000]
Bbox Number: 10
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
<|cat_start|>**figure**<|cat_end|><|box_start|>**2421 3385**<|box_end|>;<|cat_start|>**caption**<|cat_end|><|box_start|>**2345 3028**<|box_end|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>...

Task Prompt
Please perform the cwh layout generation task based on the above information. Specifically, the cwh task provides the category and size information for each bbox, and you need to predict the position coordinates for them....

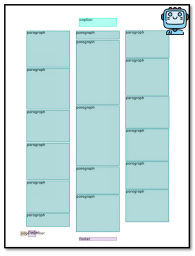


COMPLETION

Base Prompt
Document Type: "magazine"
Canvas Size: [719, 1000]
Bbox Number: 21
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
<|cat_start|>**header**<|cat_end|><|box_start|>**289 1052 2147 3034**<|box_end|><|cat_start|>**paragraph**<|cat_end|><|box_start|>**8 3 1454 2169 3054**<|box_end|>

Task Prompt
Please perform the completion layout generation task based on the above information. Specifically, the completion task provides partial bboxes, and you need to complete the remaining layout elements accordingly....

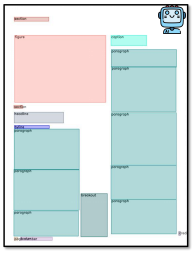


REFINEMENT

Base Prompt
Document Type: "magazine"
Canvas Size: [750, 1000]
Bbox Number: 18
Valid Categories: {"QR code", "advertisement", "author", ...}

Condition Prompt
<|cat_start|>**breakout**<|cat_end|><|box_start|>**314 1783 2108 3179**<|box_end|>;<|cat_start|>**figure**<|cat_end|><|box_start|>**40 1124 2377**

Task Prompt
Please perform the refinement layout generation task based on the above information. Specifically, the refinement task provides each bbox perturbed by noise, and you need to adjust and optimize them to improve the layout quality....



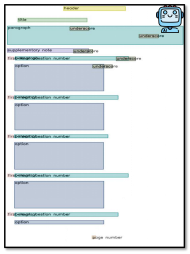
Magazine

U-COND

Base Prompt
Document Type: "exam"
Canvas Size: [562, 1000]
Bbox Number: 25
Valid Categories: {"QR code", "author", "bracket", ...}

Condition Prompt
None

Task Prompt
Please perform the unconditional layout generation task based on the above information. Specifically, the unconditional task does not provide any existing bbox information, and you need to generate the entire layout from scratch....

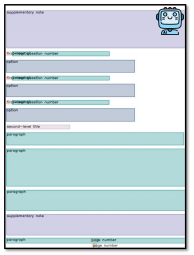


C→S+P

Base Prompt
Document Type: "exam"
Canvas Size: [681, 1000]
Bbox Number: 18
Valid Categories: {"QR code", "author", "bracket", ...}

Condition Prompt
<|cat_start|>**page number**<|cat_end|>;<|cat_start|>**supplementary note**<|cat_end|>;<|cat_start|>**first-level question number**<|cat_end|>;<|cat_start|>**paragraph**<|cat_end|>;

Task Prompt
Please perform the c layout generation task based on the above information. Specifically, the c task provides the category information for each bbox, and you need to predict both the position coordinates and size information for them....

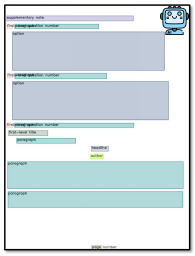


C+S→P

Base Prompt
Document Type: "exam"
Canvas Size: [695, 1000]
Bbox Number: 16
Valid Categories: {"QR code", "author", "bracket", ...}

Condition Prompt
<|cat_start|>**supplementary note**<|cat_end|><|box_start|>**2480 3021**<|box_end|>;<|cat_start|>**first-level question number**<|cat_end|><|box_start|>**2022 3019**<|box_end|>;

Task Prompt
Please perform the cwh layout generation task based on the above information. Specifically, the cwh task provides the category and size information for each bbox, and you need to predict the position coordinates for them....

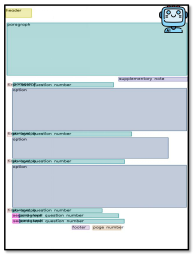


COMPLETION

Base Prompt
Document Type: "exam"
Canvas Size: [562, 1000]
Bbox Number: 20
Valid Categories: {"QR code", "author", "bracket", ...}

Condition Prompt
<|cat_start|>**header**<|cat_end|><|box_start|>**0 1012 2080 3036**<|box_end|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>**4 1070 2556 3215**<|box_end|>;<|cat_start|>**supplementary note**...

Task Prompt
Please perform the completion layout generation task based on the above information. Specifically, the completion task provides partial bboxes, and you need to complete the remaining layout elements accordingly....

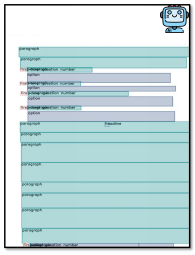


REFINEMENT

Base Prompt
Document Type: "exam"
Canvas Size: [659, 1000]
Bbox Number: 26
Valid Categories: {"QR code", "author", "bracket", ...}

Condition Prompt
<|cat_start|>**paragraph**<|cat_end|><|box_start|>**48 1174 2619 3041**<|box_end|>;<|cat_start|>**paragraph**<|cat_end|><|box_start|>**5 3 1215 2589 3044**<|box_end|>;...

Task Prompt
Please perform the refinement layout generation task based on the above information. Specifically, the refinement task provides each bbox perturbed by noise, and you need to adjust and optimize them to improve the layout quality....



Exam

Figure 18. Examples of Layouts Generated by Our OmniDocLayout-LLM on Five Generation Tasks (Magazine and Exam).

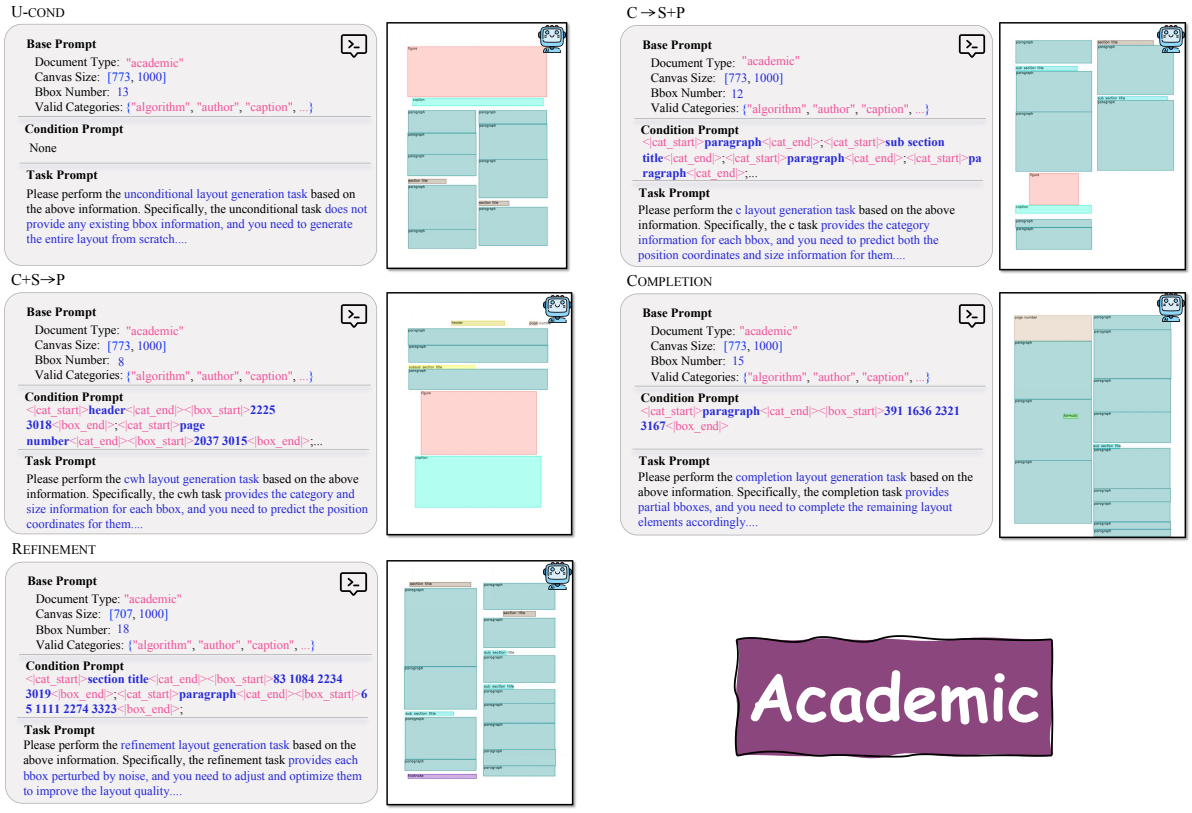


Figure 19. Examples of Layouts Generated by Our OmniDocLayout-LLM on Five Generation Tasks (Academic).