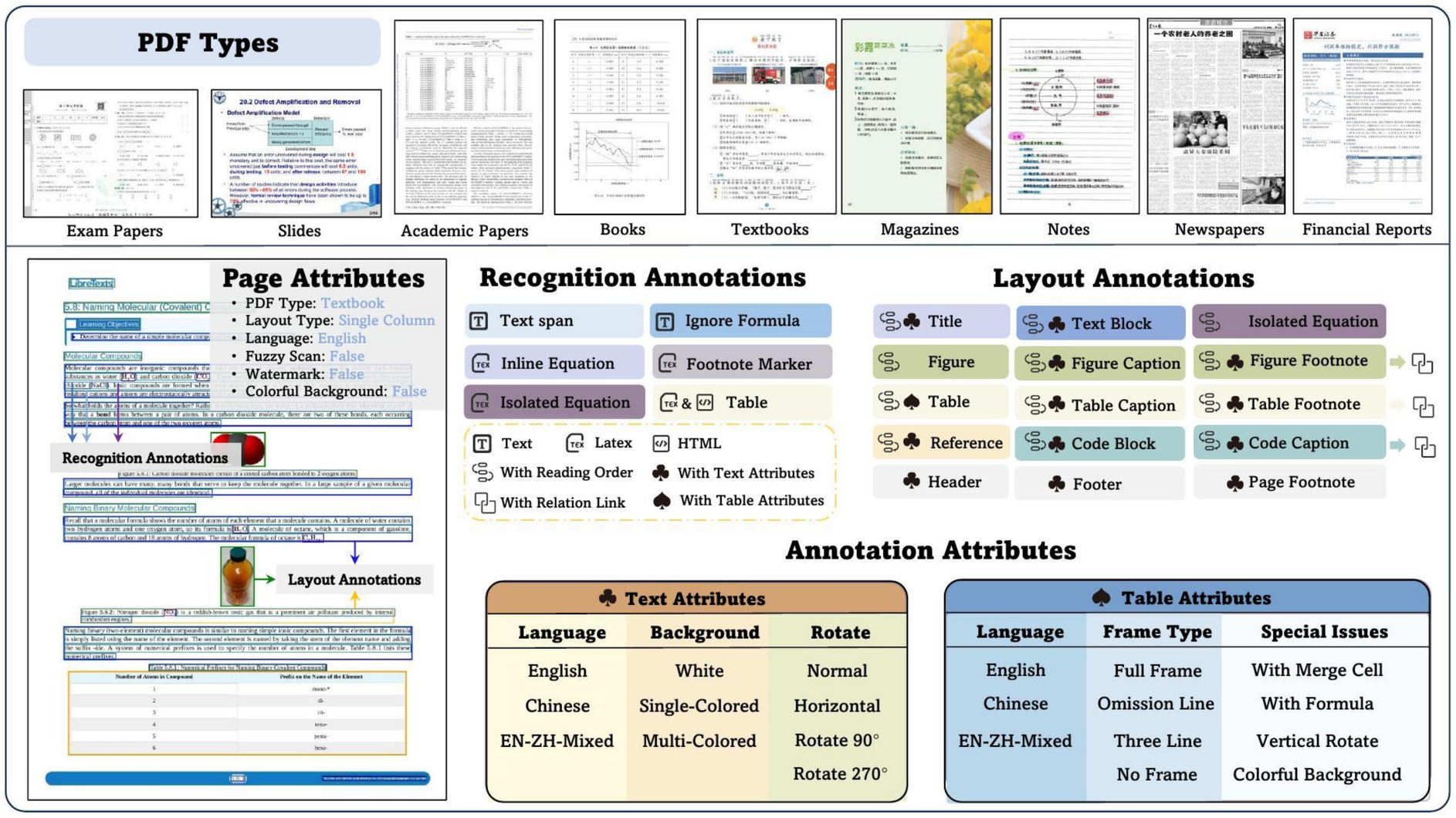

Figure 2. Overview of OmniDocBench Data Diversity. The benchmark includes 9 diverse PDF document types. It supports rich annotation types, including layout annotations (e.g., title, table, figure) and recognition annotations (e.g., text spans, equations, tables). Each page is annotated with 6 page-level attributes (e.g., PDF type, layout type), along with fine-grained 3 text attributes (e.g., language) and 6 tables attributes (Items under "*special issues*" are treated as individual binary attributes (yes/no)), enabling detailed and robust evaluation.

developed to target specific sub-tasks. For end-to-end evaluation, works like Nougat [7] and GOT-OCR [45] provide relatively small validation sets and assess predictions using page-level metrics such as Edit Distance [21].

However, these benchmarks present several key limitations: 1) **Limited document diversity:** Existing datasets primarily focus on academic papers, overlooking other real-world document types such as textbooks, exams, financial reports, and newspapers; 2) **Inconsistent evaluation metrics:** Current benchmarks rely heavily on generic text similarity metrics (e.g., Edit Distance [21] and BLEU [33]), which fail to fairly assess the accuracy of formulas and tables in LaTeX or HTML formats that allow for diverse syntactic expressions; and 3) **Lack of fine-grained evaluation:** Most evaluations report only an overall score, lacking insights into specific weaknesses, such as element-level score (e.g., text vs. formula) or per document-type performance (e.g., magazine or notes).

To address these limitations, we introduce **OmniDocBench**, a new benchmark designed to provide a rigorous and comprehensive evaluation for document parsing models across both pipeline-based and end-to-end paradigms. In summary, our benchmark introduces the following key contributions:

- **High-quality, diverse evaluation set:** We include pages from 9 distinct document types, ranging from textbooks to newspapers, annotated using a combination of automated tools, manual verification, and expert review.
- **Flexible, multi-dimensional evaluation:** We support comprehensive evaluation at three levels—end-to-end, task-specific, and attribute-based. End-to-end evaluation measures the overall quality of full-page parsing results. Task-specific evaluation allows users to assess individual components such as layout detection, OCR, table recognition, or formula parsing. Attribute-based evaluation provides fine-grained analysis across 9 document types, 6 page-level attributes and 9 bbox-level attributes.
- **Comprehensive benchmarking of state-of-the-art methods:** We systematically evaluate a suite of representative document parsing systems, including both pipeline-based tools and VLMs, providing the most comprehensive comparison and identifying performance bottlenecks across document types and content structures.

2. Related Work

2.1. Pipeline-based Document Content Extraction

Pipeline-based methods treat the document content extraction task as a collection of single modules, such as document layout detection [13, 17, 36, 53], optical character recognition [15, 23, 30, 38, 43], formula recognition [6, 27, 40, 51], and table recognition [16, 18, 23]. In