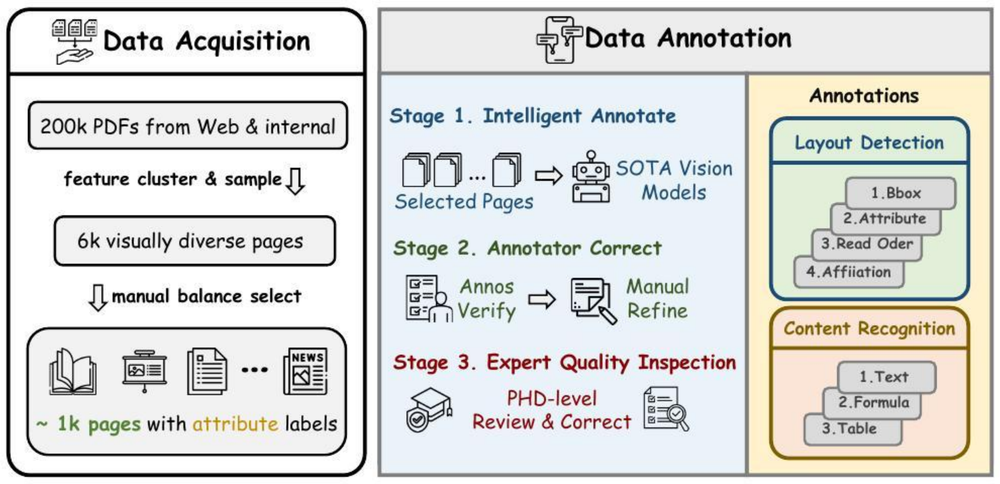

Figure 3. Overview of the OmniDocBench dataset construction.

3. OmniDocBench Dataset

Constructing a diverse and comprehensive document parsing benchmark with precise annotations is a significant challenge. As illustrated in Figure 3, we have designed a systematic and professional annotation framework for OmniDocBench, encompassing data acquisition, intelligent pre-annotation, and manual refinement. This ensures that OmniDocBench possesses the following key attributes:

- **Page Diversity**. We sourced document pages from a variety of origins to ensure a wide range of document types.
- **Comprehensive Annotation**. We meticulously annotated all elements on the pages, including bounding boxes, specific contents, and various potential attributes.
- **Annotation Accuracy**. By integrating semi-automated annotation processes, annotator corrections, and expert quality checks, we ensure the reliability of all annotations.

The following sections detail the data acquisition process, the annotation methodology, and a statistical analysis of the final annotated dataset.

3.1. Data Acquisition

During the data acquisition phase, we sourced document pages from diverse origins and used clustering algorithms to initially select visually diverse pages, followed by manual annotation of page attributes to finalize the OmniDocBench pages. Specifically, we collected over 200,000 initial PDF documents from Common Crawl, Google, Baidu search engines, and internal data. Subsequently, we extracted visual features from these document pages using ResNet-50 and performed clustering using Faiss [1], sampling 6,000 visually diverse pages from 10 cluster centers. Finally, annotators provided page-level attribute annotations, including page type, layout type, and language type, and further balanced the selection to 981 samples for the final dataset. The OmniDocBench dataset includes pages from nine distinct types, multiple layout categories, and various attribute annotations, covering a wide range of real-world scenarios.

[1] https://github.com/facebookresearch/faiss

3.2. Data Annotation

To ensure the comprehensiveness of OmniDocBench's annotations, we conducted detailed annotations for layout detection and content recognition.

3.2.1. Annotation Types

Layout Detection Annotations: Unlike typical layout detection tasks, OmniDocBench includes four comprehensive types of annotations: (1) Layout Bounding Box Annotations: Positioanl information for 19 distinct region categories such as titles, text paragraphs, tables, and images. (2) Layout Attribute Annotations: Detailed attribute annotations for detected boxes, including 3 text box attribute categories, 6 table attribute categories, 9 bbox-level attribute labels in total. (3) Reading Order Annotations: Annotating the reading sequence of detected boxes. (4) Affiliation Annotations: For images, tables, formulas, and code blocks, we annotate captions and titles to distinguish them from main text. Similarly, for cross-page paragraphs, we annotate affiliation relationships.

Content Recognition Annotations: Based on the content type within each region, we conduct the following three types of annotations: (1) Text Annotations: Pure text annotations for titles, text paragraphs, and other plain text content. (2) Formula Annotations: LaTeX format annotations for inline formulas, display formulas, and subscripts. (3) Table Annotations: Providing both HTML and LaTeX annotations for table data.

3.2.2. Annotation Process

For these annotation tasks on diverse pages, we design a standardized process to ensure quality and efficiency, comprising intelligent automatic annotation, annotator correction, and expert quality inspection.

Automatic Annotation. Manually annotating entire documents is time-consuming and costly. To enhance efficiency, we employ state-of-the-art detection and recognition models for pre-annotation of layout detection and content recognition. Specifically, we use fine-tuned LayoutLMv3 [17] for layout detection annotations and PaddleOCR [23], UniMERNet [40], and GPT-4o [2] for text, formula, and table annotations, respectively.

Annotator Correction. After the layout detection phase, annotators refine the detection boxes and enhance annotations with reading order and affiliation details. Each character is verified to ensure accuracy in content recognition. For complex annotations of tables and formulas, requiring LaTeX and HTML formats, annotators use tools like Tables Generator [2] and latexlive [3] for verification and correction.

Expert Quality Inspection. Despite thorough annotator corrections, the complexity of formulas and tables may re-

[2] https://www.tablesgenerator.com/
[3] https://www.latexlive.com/