MoDora: A Multimodal Document AI Assistant Harness

Yukai Wu
Shanghai Jiao Tong University
wuyukai99@gmail.com

Bangrui Xu
Shanghai Jiao Tong University
dreameternal@sjtu.edu.cn

Shaoli Yu
Shanghai Jiao Tong University
geniustanker@sjtu.edu.cn

Zirui Tang
Shanghai Jiao Tong University
tangzirui@sjtu.edu.cn

Xuanhe Zhou
Shanghai Jiao Tong University
zhouxuanhe@sjtu.edu.cn

Yeye He
Microsoft Research
yeyehe@microsoft.com

Bin Wang
Shanghai AI Lab
wangbin@pjlab.org.cn

Conghui He
Shanghai AI Lab
heconghui@pjlab.org.cn

Guoliang Li
Tsinghua University
liguoliang@tsinghua.edu.cn

Fan Wu
Shanghai Jiao Tong University
fwu@cs.sjtu.edu.cn

ABSTRACT

General-purpose AI document assistants (e.g., NotebookLM) increasingly play an important role and are widely adopted across diverse domains. However, they consistently struggle on complicated multimodal documents such as financial reports and scientific papers, where hierarchical structures, complex layouts, and interleaved visual elements carry essential semantics. A key reason is that these assistants lack native multimodal data management capabilities: they typically linearize documents into flat text chunks or page images, discarding the logical hierarchy and cross-modal context that are critical for faithful evidence retrieval and reasoning. To address these limitations, we present MoDora, an interactive, tree-structured multimodal document analysis agent harness. Designed to make the document analysis process transparent and controllable, MoDora introduces an end-to-end user experience through three core functionalities: (1) an automated document ingestion engine that seamlessly transforms unstructured PDFs into a layout-aware component tree, preserving both textual hierarchy and visual elements; (2) an interactive structure visualizer that allows users to intuitively inspect the extracted document hierarchy and refine structural relations via drag-and-drop or natural language commands; and (3) a verifiable multimodal QA interface that supports complex, cross-modal queries with precise bounding-box-level grounding back to the original PDF regions, enabling users to effortlessly trace and verify the underlying evidence. Experimental results on the MMDA benchmark show that MoDora achieves an AIC-Acc of 71.1%, outperforming baselines by over 14%.

PVLDB Reference Format:
Yukai Wu, Bangrui Xu, Shaoli Yu, Zirui Tang, Xuanhe Zhou, Yeye He, Bin Wang, Conghui He, Guoliang Li, and Fan Wu. MoDora: A Multimodal Document AI Assistant Harness. PVLDB, 19(12): 4794 - 4797, 2026.
doi:10.14778/3827998.3828124

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:10.14778/3827998.3828124

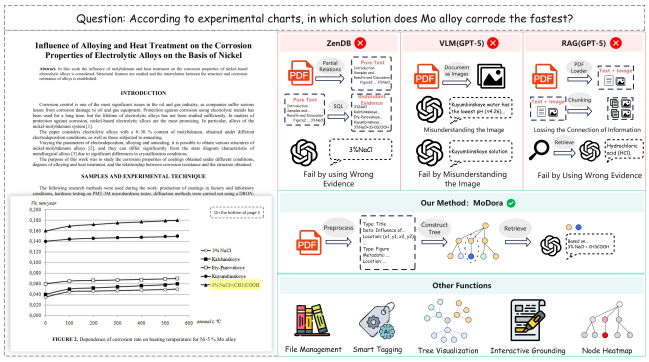

Figure 1: Overview of the MoDora Demonstration.

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/OpenDataBox/MoDora.

1 INTRODUCTION

AI-powered document tools have become essential for helping human workers to extract actionable insights from raw documents. Widely used products such as Google NotebookLM and ChatGPT with file upload have made document-grounded analysis increasingly accessible to end users. Meanwhile, research systems such as DocAgent and M3DocRAG further explore long-context multimodal document understanding and retrieval, while DocOwl studies OCR-free multimodal document understanding [3, 10, 15]. However, these general-purpose assistants consistently underperform on multimodal documents (e.g., those with interleave hierarchical text, tables, charts, and images in complex layouts [12]). For instance, when asked about a specific data point in a scientific paper, a general assistant may hallucinate values because it flattens the table into plain text, misidentifies sidebar content as main body, or fails to link a chart to its describing paragraph [13].

We argue that the root cause lies in the absence of systematic multimodal data management capabilities within these assistants. Existing approaches either process documents end-to-end with models that jointly reason over visual, textual, and layout signals [11, 15],

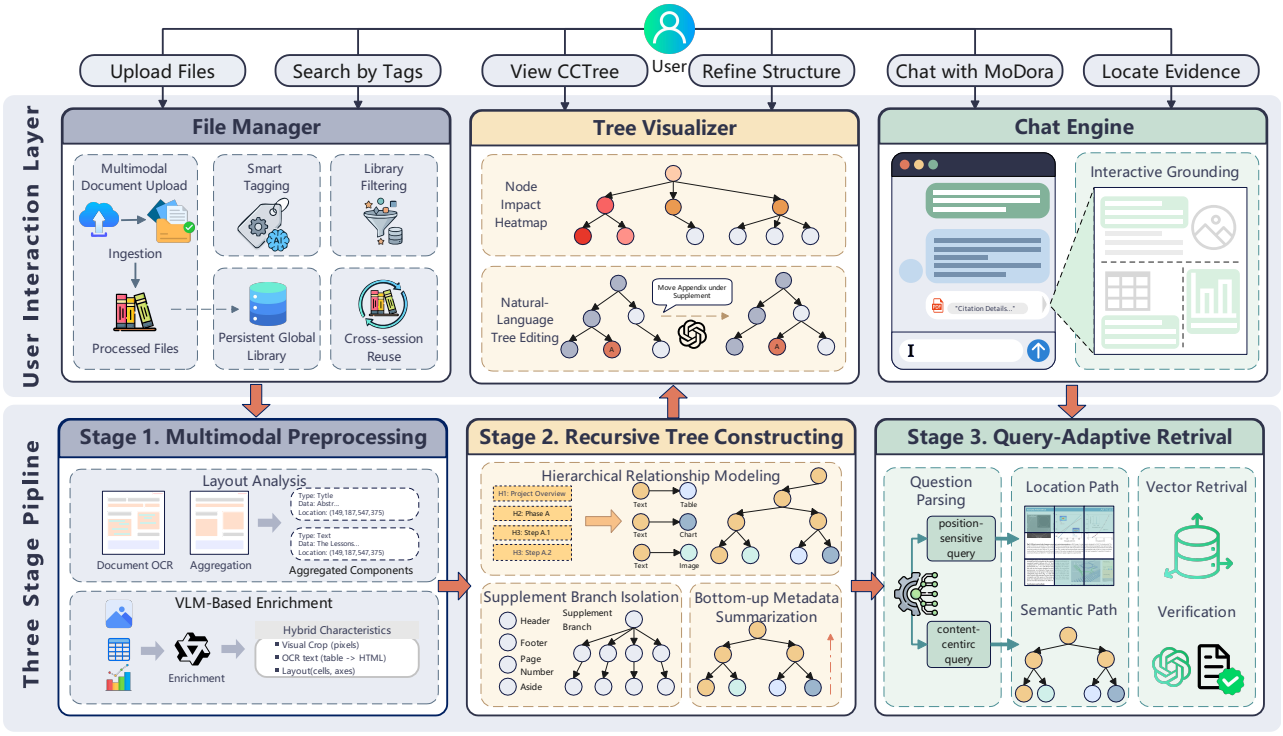

Figure 2: System Architecture of MoDora.

or index them for retrieval at passage or page granularity [3, 5]. Neither paradigm captures the rich, nested structure inherent in real-world documents. Recent data management research has begun to address document structure. For example, ZenDB [6] extracts semantic heading trees for SQL-style analytics, and EVAPORATE [1] generates structured views from data lakes. However, these systems remain largely text-only and cannot handle visual elements such as charts and tables that carry critical information. As a result, three fundamental data management challenges remain open. *(1) Fragmented Element Aggregation.* Low-level parsing tools, such as OCR engines, produce fragmented elements that are stripped of their original semantic context, making it difficult to form coherent, self-contained units for downstream analysis [13]. *(2) Hierarchical and Layout-Aware Representation.* Documents exhibit nested hierarchies, for example, chapters containing sections and sections containing figures, as well as layout-specific distinctions, such as sidebars versus the main body, which flat-chunk or page-level representations fail to capture [8, 14]. *(3) Cross-Modal Evidence Retrieval and Alignment.* Answering complex questions often requires retrieving and aligning evidence scattered across multiple regions, pages, and modalities, for example, linking a descriptive paragraph to a table cell located elsewhere in the document [3, 13].

To address these challenges, we develop MoDora, an agentic tree-based multimodal document management and analysis system. As shown in Figure 1, MoDora represents unstructured PDFs as hierarchical Component-Correlation Trees (CCTrees), which capture both document hierarchy and visual semantics. Built on this abstraction, MoDora provides an end-to-end human-in-the-loop analysis pipeline that supports structure-aware document ingestion, interactive CCTree inspection and refinement, grounded multimodal retrieval, and observable answer verification.

Compared with our research paper [13], this demonstration focuses on the system-level integration and user-facing interaction design of MoDora. In particular, the demo exposes interfaces for document collection management, smart tagging, tree visualization, interactive grounding, and node-level retrieval heatmaps. These interfaces allow users to organize document collections, inspect and revise extracted structures, verify retrieved evidence against the original PDF, and improve retrieval accuracy by manually editing tree structures or refining them through natural-language instructions. We have implemented MoDora and evaluated it on the MMDA benchmark. Experimental results show that MoDora achieves an AIC-Acc of 71.1%, outperforming eight representative baselines [2, 3, 5, 6, 9–11, 15] by more than 14%.

In summary, this demonstration makes the following contributions: (1) a CCTree-based multimodal document management system that preserves structural and visual semantics for document analysis; (2) an interactive human-in-the-loop pipeline for ingestion, tree construction, multimodal QA, evidence grounding, structure editing, and retrieval inspection; and (3) a user-facing demonstration and evaluation showing how CCTree-based analysis supports transparent and verifiable multimodal document understanding.

2 SYSTEM ARCHITECTURE

The architecture of MoDora is designed to transform unstructured, multimodal documents into a structured knowledge hierarchy while

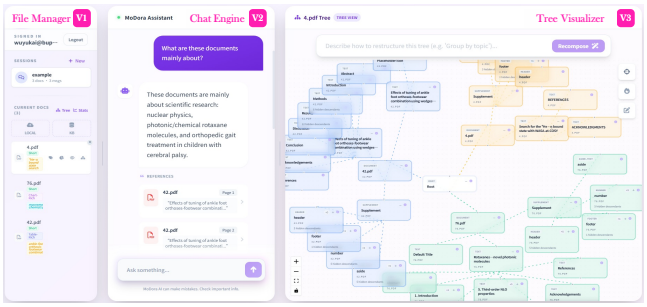

Figure 3: User Interface of MoDora

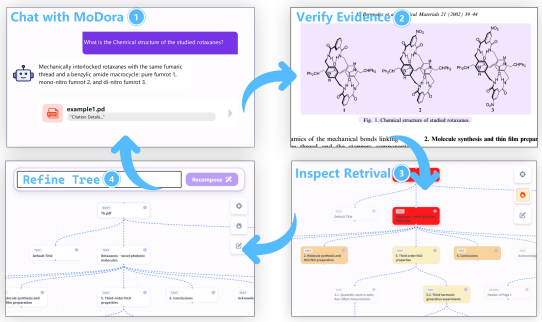

Figure 4: End-to-end interaction loop in MoDora.

providing an interactive and verifiable user experience. As illustrated in Figure 2, the system consists of two layers: a User Interaction Layer and a Three-Stage Document Analysis pipeline.

User Interaction Layer. MoDora provides an integrated interface with three user-facing modules: a File Manager for document upload and cross-session reuse, a Chat Engine for multimodal question answering with interactive grounding, and a Tree Visualizer for CCTree inspection, node heatmap visualization, and natural-language structure refinement.

Three-Stage Document Analysis. Behind the interactive frontend, MoDora employs a robust three-stage pipeline to process documents and answer queries.

- ***Stage 1: Component-Level Document Preprocessing.*** MoDora first converts raw PDF and OCR outputs into structure-aware multimodal components, preserving both textual content and key visual information needed for later retrieval and grounding.
- ***Stage 2: Hierarchical Tree Construction.*** The resulting components are then organized into a hierarchical CCTree, which preserves document structure, captures layout distinctions, and provides a stable basis for inspection, retrieval, and downstream interaction.
- ***Stage 3: Grounded Retrieval and Answering.*** Finally, MoDora performs multimodal retrieval and answer generation over the CCTree. The evidence is grounded back to the PDF and recorded as provenance, enabling users to inspect supporting regions, verify system responses, and iteratively refine document understanding.

3 END-TO-END EXPERIENCE

This section demonstrates how users turn unstructured multimodal documents into verifiable knowledge assets. Figure 3 shows the layout: V1 is the File Manager, V2 is the Chat Engine, and V3 is the Tree Visualizer. Figure 4 illustrates the loop connecting answering, verification, retrieval inspection, and structure refinement.

Document Ingestion and Automated Modeling. Users begin in V1, the file and session manager, where they upload documents, browse processed files, and switch sessions. Once a document is submitted, MoDora converts it into a structured CCTree displayed in the interface. A persistent global library enables cross-session reuse, and summary statistics plus semantic tags help users understand document composition before deeper inspection.

Transparent Structure Inspection and Editing. V3 presents sections, paragraphs, tables, and figures as an editable CCTree, making missing nodes, incorrect parent-child relations, or misplaced visual components directly observable. Users refine the tree by drag-and-drop or natural-language instructions such as "Move '3.2 Results' under Section III", after which the system updates the CCTree and runs later retrieval over the revised structure.

Verifiable Multimodal Question Answering. V2 hosts the conversational assistant. For each natural-language query, MoDora classifies the query type (position-sensitive vs. content-centric), selects the CCTree retrieval path, and returns answers with inline citations linked to CCTree nodes. Clicking a citation activates interactive grounding in the original PDF, highlighting the exact bounding box of the relevant paragraph, table, or chart while preserving headers, sidebars, figures, and layout cues for transparent verification.

Retrieval Provenance and Iterative Refinement. Figure 4 shows the closed human-in-the-loop cycle. After receiving an answer (step ①), users verify cited evidence through PDF grounding (step ②) and inspect retrieval behavior via the Node Heatmap (step ③), whose color intensity shows which CCTree nodes contributed most strongly. If evidence is incomplete or mis-structured, users revise the tree editor (step ④) and reissue the query, turning MoDora into an iterative document-understanding environment.

4 EVALUATION

In our evaluation, document preprocessing uses PPStructureV3 [4] for element extraction; Qwen3-VL-8B [7] and Qwen3-Embedding-8B are separately deployed on two NVIDIA GeForce RTX 4090 GPUs (24GB each) for template-based enrichment/backward verification and embedding retrieval. GPT-5 serves as the backbone LLM for format-aware hierarchy detection and answer generation. As a system, these model components are modular services and can also be readily accessed through hosted APIs instead of local GPUs.

4.1 Quantitative Comparison

We first conduct quantitative evaluation on the MMDA benchmark [13], which contains 537 multimodal documents spanning scientific reports, financial statements, and technical manuals, with 1065 questions covering five types: text-only, hybrid-data, structural-hierarchy, location-based and metadata-related.

As shown in Figure 5, MoDora achieves 71.1% AIC-Acc, outperforming the strongest baseline (DocAgent, 57.0%) by 14.1 percentage points. The gap is clear against methods without explicit structural modeling: GPT-5 (46.5%) treats documents as flat page images,

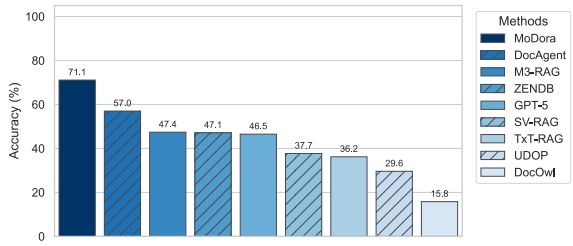

Figure 5: Experimental results on the MMDA dataset.

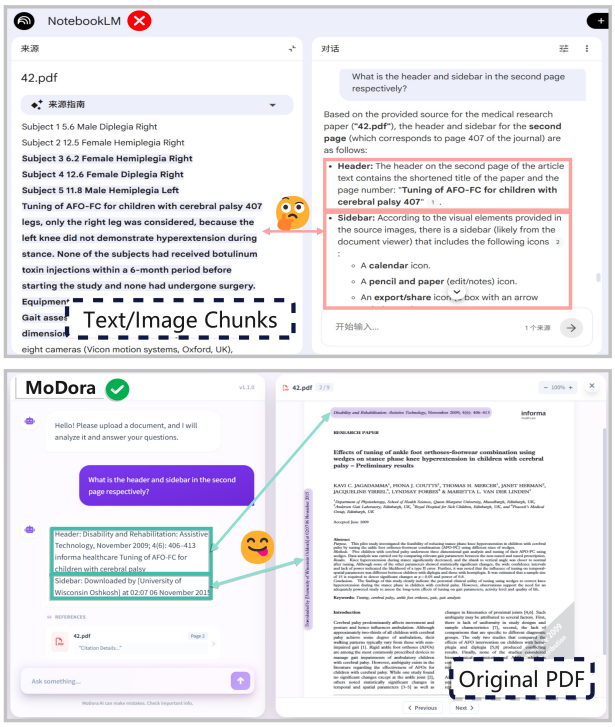

Figure 6: User-facing comparison of evidence presentation.

text-only RAG methods such as TxT-RAG (36.2%) and SV-RAG (37.7%) discard visual elements, and end-to-end models such as UDOP (29.6%) and DocOwl (15.8%) face sequence-length and coarse-encoding limits. MoDora preserves hierarchy and cross-modal component semantics for more precise evidence retrieval.

4.2 User-Facing Comparison

To contextualize MoDora's capabilities beyond benchmark accuracy, we compare it with Google NotebookLM, a representative general-purpose document assistant, on the same multimodal document. Figure 6 presents a side-by-side comparison of how each system presents evidence to users.

Evidence Presentation. When asked "What is the header and sidebar in the second page respectively?", NotebookLM returns a text-based answer assembled from extracted text chunks (upper panel of Figure 6). The source panel lists document segments as flat text or image snippets without spatial context, making it difficult for users to verify whether the identified "header" and "sidebar" correspond to the correct visual regions. In contrast, MoDora (lower panel of Figure 6) provides the answer with interactive grounding: clicking the citation highlights the precise bounding box in the original PDF, showing the header region at the top and the sidebar region on the left of page 2. This grounding preserves the original layout, font formatting, and surrounding visual context, enabling users to verify the answer at a glance.

Structure-Aware. Beyond answer presentation, MoDora exposes an explicit document structure through the CCTree. This structure-aware view helps separate main-body content from supplementary regions such as headers, footers, and sidebars, enabling more precise evidence tracing for document-structure-related queries (e.g., "What are the section titles?"). In contrast, general-purpose assistants typically expose retrieved evidence as text/image snippets rather than an editable structural view of the document. As a result, MoDora provides a more transparent user experience for inspecting, verifying, and refining document understanding.

ACKNOWLEDGMENTS

Xuanhe Zhou is the corresponding author. This work was supported in part by National Key R&D Program of China (No. 2023YFB4502400), China NSF grants (No. 62441236, 62372296, 62432007, U25A6024, U25A20437, 62525202, 62232009, 62502304), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM103), ByteDance, Tencent, Ant Group through CCF-Ant Research Fund, Tencent Rhino Bird Key Research Project, Huawei Innovation Research Programs, Shenzhen Project (CJGJZD20230724093403007), Zhongguancun Lab, Beijing National Research Center for Information Science and Technology (BNRist), and Shanghai Jiao Tong University AI for Engineering Initiative.

REFERENCES

[1] S. Arora, B. Yang, S. Eyuboglu, et al. 2023. Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes. VLDB 2023.
[2] J. Chen, R. Zhang, Y. Zhou, et al. 2025. SV-RAG: LoRA-Contextualizing Adaptation of MLLMs for Long Document Understanding. ICLR 2025.
[3] J. Cho, D. Mahata, O. Irsoy, et al. 2024. M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding. arXiv:2411.04952.
[4] C. Cui, T. Sun, M. Lin, et al. 2025. PaddleOCR 3.0 Technical Report. arXiv:2507.05595.
[5] P. Lewis, E. Perez, A. Piktus, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.
[6] Y. Lin, M. Hulsebos, R. Ma, et al. 2024. Towards accurate and efficient document analytics with large language models. arXiv:2405.04674.
[7] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL]
[8] J. Saad-Falcon, J. Barrow, A. Siu, et al. 2024. PDFTriage: Question Answering over Long, Structured Documents. EMNLP Industry 2024.
[9] A. Singh, A. Fry, A. Perelman, et al. 2026. OpenAI GPT-5 System Card. arXiv:2601.03267.
[10] L. Sun, L. He, S. Jia, et al. 2025. DocAgent: An Agentic Framework for Multi-Modal Long-Context Document Understanding. EMNLP 2025.
[11] Z. Tang, Z. Peng, S. Geng, et al. 2023. Unifying Vision, Text, and Layout for Universal Document Processing. CVPR 2023.
[12] D. Wang, N. Raman, M. Sibue, et al. 2024. DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding. ACL 2024.
[13] B. Xu, Q. Yao, Z. Tang, et al. 2026. MoDora: Tree-Based Semi-Structured Document Analysis System. SIGMOD 2026.
[14] C. Yang, D.-K. Vu, M.-T. Nguyen, et al. 2025. SuperRAG: Beyond RAG with Layout-Aware Graph Modeling. NAACL Industry 2025.
[15] J. Ye, A. Hu, H. Xu, et al. 2023. mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding. arXiv:2307.02499.