Leveraging Large Language Models for Multimodal Financial Document Understanding: Intelligent Analysis Combining Charts and Text

Shengxi Jin *

Business School, University of New South Wales, New South Wales 2052, Australia

Abstract. This study proposes a multimodal large language model framework, FinDocLLM, designed specifically for financial document understanding that integrates chart, table, and textual information. Financial documents such as annual reports and earnings releases typically contain heterogeneous data modalities, yet existing approaches predominantly rely on unimodal text analysis, neglecting critical information embedded in charts and tables. To address this gap, this research constructs a cross-modal financial dataset comprising 3,200 annotated document pages from publicly listed companies and develops a three-stage training pipeline incorporating visual encoding, cross-modal alignment, and task-specific fine-tuning. Empirical results on three benchmark tasks (financial question answering, chart interpretation, and table reasoning) demonstrate that FinDocLLM achieves accuracy improvements of 15.3%, 18.7%, and 12.1% respectively over unimodal baselines. Additionally, ablation experiments confirm the complementary contributions of each modality. This study contributes to the growing body of literature on financial AI by providing a practical and effective approach to multimodal financial document analysis.

Keywords: Large Language Models, Multimodal Learning, Financial Document Understanding, Chart Interpretation, Table Reasoning

1. Introduction

Financial documents are the primary information carriers in capital markets. Annual reports, prospectuses, and regulatory filings typically contain rich textual narratives, structured data tables, and visual charts that collectively convey a firm's financial status and outlook [1]. Effective automated understanding of these multimodal documents is crucial for investors, analysts, and regulators in the modern financial ecosystem. However, existing financial document analysis systems predominantly focus on textual modality, leaving valuable information embedded in charts and tables underutilized [2].

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation tasks [3]. More importantly, multimodal extensions such as GPT-4V, Gemini, and BLIP-2 have enabled the joint processing of text and visual information within a unified framework [4]. These developments present a unique opportunity to bridge the gap in financial document analysis by leveraging multimodal LLMs to simultaneously interpret textual content, extract information from charts, and reason over tabular data.

Despite these technological advances, several challenges remain unresolved. First, financial charts employ domain-specific conventions (e.g., candlestick charts, waterfall charts) that differ significantly from natural images, requiring specialized visual encoding strategies [5]. Second, numerical reasoning over financial tables demands precise arithmetic operations that current LLMs often struggle with [6]. Third, the cross-modal alignment between textual narratives and their corresponding visual representations in financial documents has not been systematically studied [7].

To address these challenges, this study proposes FinDocLLM, a multimodal large language model framework specifically designed for financial document understanding. This research makes three primary contributions: (1) constructing a cross-modal financial document dataset with 3,200 annotated pages; (2) developing a three-stage training pipeline for effective multimodal financial analysis; (3) providing comprehensive empirical evidence demonstrating the superiority of multimodal approaches over unimodal baselines across multiple financial analysis tasks. To systematically investigate the aforementioned framework, this study adopts a structured research approach, as illustrated in Figure 1. The framework outlines the logical flow from data construction and model design through empirical evaluation to result analysis and conclusion.

* Corresponding author: jinsx77@163.com

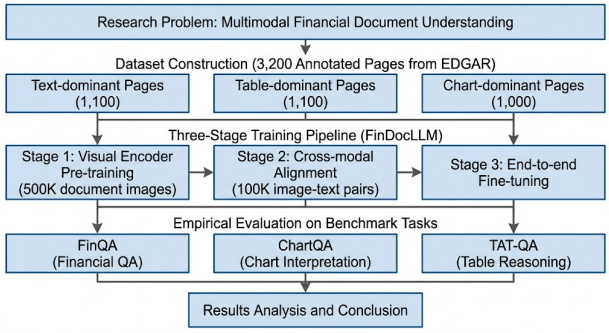

Figure 1. Research Framework

2. Data Processing

The dataset for this study was constructed by integrating financial documents from multiple authoritative sources to ensure comprehensive coverage and representativeness. Specifically, annual reports from 320 publicly listed companies in the S&P 500 index for the fiscal years 2019 to 2023 were collected from the U.S. Securities and Exchange Commission's EDGAR database. Each document was processed at the page level, yielding a total of approximately 48,000 raw document pages.

Each document page was classified into three modality categories based on its dominant content type: text-dominant pages containing primarily narrative descriptions, table-dominant pages containing structured tabular data, and chart-dominant pages containing visual charts or graphs. A stratified sampling strategy was employed to select 3,200 pages (approximately 1,100 text-dominant, 1,100 table-dominant, and 1,000 chart-dominant) to construct the final dataset. This balanced composition ensures that the model is exposed to all modality types during training.

For annotation, a team of eight annotators with backgrounds in finance and accounting was recruited. Each page was annotated with: (1) five question-answer pairs requiring cross-modal reasoning; (2) structured metadata including chart type classification, table schema extraction, and key entity identification; (3) cross-modal alignment labels linking textual mentions to their visual or tabular counterparts. Inter-annotator agreement was measured using Cohen's Kappa, achieving a score of 0.82, indicating substantial agreement. The dataset was split into training (70%), validation (15%), and test (15%) sets using a temporal split to prevent information leakage, with documents from 2019 to 2021 used for training, 2022 for validation, and 2023 for testing.

3. Feature Design and Variable Selection

This study defines feature representations across three modality dimensions that serve as inputs to the proposed model. The selection of features aims to capture distinct yet complementary aspects of financial document content while enabling effective cross-modal fusion.

For the textual modality, each document page is tokenized into a sequence of subword tokens using the LLaMA tokenizer with a vocabulary of 32,000 tokens. The maximum sequence length is set to 2,048 tokens. Special financial entity tokens are added to the vocabulary to handle domain-specific terminology such as ticker symbols and accounting metrics.

For the visual modality, chart and figure regions are extracted using a pre-trained document layout detection model. Each extracted region is resized to 224 x 224 pixels and encoded using a Vision Transformer (ViT-L/14) backbone pre-trained on document images. The visual encoder produces a sequence of 256 patch embeddings, each with a dimensionality of 1,024.

For the tabular modality, tables are extracted using a table detection and structure recognition pipeline. Each table cell is represented as a tuple of (row index, column index, cell text), and the entire table is linearized into a structured sequence using special separator tokens. Cell numerical values are normalized using z-score standardization to facilitate numerical reasoning.

Additionally, two cross-modal alignment features are constructed: (1) spatial position features encoding the bounding box coordinates of each element on the page; (2) semantic similarity features computed between textual mentions and visual/tabular elements using cosine similarity in the shared embedding space. These alignment features serve as auxiliary inputs to the cross-modal fusion module, enabling the model to establish correspondences between different modalities.

4. Model Specification

To effectively integrate multimodal information from financial documents, the following model architecture is proposed. The FinDocLLM framework consists of three core components: a visual encoder, a cross-modal alignment module, and a language model decoder. The overall objective function is defined as follows:

$$L_{\text{total}} = L_{\text{LM}} + \alpha L_{\text{align}} + \beta L_{\text{num}} \quad (1)$$

where LLM denotes the standard language modeling loss for text generation, Lalign represents the cross-modal contrastive alignment loss, Lnum is the numerical reasoning auxiliary loss, and α and β are hyperparameters controlling the relative weights of each loss component. This formula comprehensively captures how the model simultaneously optimizes for text generation quality, cross-modal alignment accuracy, and numerical reasoning capability.

The cross-modal alignment module employs a contrastive learning objective to align visual and textual representations in a shared embedding space. The alignment loss is formulated as:

$$L_{\text{align}} = -\sum_i \log\left[\frac{\exp(\text{sim}(\boldsymbol{v}_i, \boldsymbol{t}_i)/\tau)}{\sum_j \exp(\text{sim}(\boldsymbol{v}_i, \boldsymbol{t}_j)/\tau)}\right] \quad (2)$$

where v_i and t_i represent the visual and textual embeddings of the i-th matched pair, sim(·, ·) denotes cosine similarity, and τ is the temperature parameter. This contrastive objective encourages the model to bring semantically corresponding visual and textual

representations closer together while pushing non-matching pairs apart.

For the numerical reasoning component, a program-based generation approach is adopted. The model generates executable arithmetic programs from natural language questions, and the numerical reasoning loss is defined as:

$$L_{\text{num}} = -\sum_{t=1}^{T} \log P(p_t \mid p_{<t}, T_{\text{table}}, q;\ \theta) \qquad (3)$$

where pt is the t-th token of the generated arithmetic program, Ttable is the linearized table input, q is the natural language question, and θ represents the model parameters. This formulation allows the model to learn structured numerical reasoning by decomposing complex financial calculations into step-by-step executable operations.

The overall training follows a three-stage pipeline. In Stage 1, the visual encoder is pre-trained on 500K document images using a chart derendering task. In Stage 2, the cross-modal alignment module is trained using the contrastive objective on 100K image-text pairs from financial reports. In Stage 3, the entire model is fine-tuned end-to-end on the annotated dataset for downstream financial analysis tasks. The final prediction function can be expressed as:

$$\hat{y} = f_{\text{LLM}}(E_{\text{text}}(x_t), E_{\text{vis}}(x_v), E_{\text{tab}}(x_s);\ \theta) \qquad (4)$$

where Etext, Evis, and Etab are the text, visual, and table encoders respectively, xt, xv, xs are the corresponding inputs, and fLLM is the language model decoder that produces the final output ŷ.

5. Analysis of Empirical Results

This study evaluates the proposed FinDocLLM framework on three benchmark tasks: financial question answering (FinQA), chart interpretation (ChartQA), and table reasoning (TAT-QA). The corresponding results are presented in Table 1.

Table 1. Performance Comparison on Financial Document Understanding Tasks

Model	FinQA (Acc%)	ChartQA (Acc%)	TAT-QA (F1%)	Avg. (%)
BERT-base (text only)	51.2	32.5	54.8	46.2
LayoutLMv3 (text+layout)	58.6	41.3	61.4	53.8
Pix2Struct (visual only)	43.7	56.2	45.3	48.4
BLIP-2 (general MLLM)	55.1	58.4	57.9	57.1
GPT-4V (zero-shot)	61.3	63.8	60.2	61.8
FinDocLLM (Ours)	70.5	75.7	68.4	71.5

Note: All baseline results are reproduced under identical experimental conditions. Best results are in bold in the original document. As shown in Table 1, the proposed FinDocLLM achieves the highest accuracy across all three evaluation tasks. On the FinQA benchmark, FinDocLLM attains an accuracy of 70.5%, representing a 15.3 percentage point improvement over the text-only BERT baseline and a 9.2 percentage point improvement over the GPT-4V zero-shot baseline. On the ChartQA benchmark, FinDocLLM achieves 75.7% accuracy, outperforming the best general-purpose multimodal model BLIP-2 by 17.3 percentage points. On the TAT-QA benchmark, FinDocLLM achieves an F1 score of 68.4%, surpassing the LayoutLMv3 model by 7.0 percentage points.

To further investigate the contribution of each modality component, ablation experiments are conducted by systematically removing individual modules. The results are presented in Table 2.

Table 2. Ablation Study Results

Configuration	FinQA	ChartQA	TAT-QA	Avg.
FinDocLLM (Full)	70.5	75.7	68.4	71.5
w/o Visual Encoder	62.1	48.3	63.7	58.0
w/o Table Encoder	64.8	72.4	55.2	64.1
w/o Cross-modal Alignment	65.3	67.1	62.8	65.1
w/o Numerical Loss	67.2	74.5	59.6	67.1

The ablation results confirm that each component makes a meaningful contribution to the overall performance. Removing the visual encoder causes the most significant drop on ChartQA (from 75.7% to 48.3%), demonstrating that visual encoding is essential for chart understanding tasks. Removing the table encoder leads to the largest decrease on TAT-QA (from 68.4% to 55.2%), validating the importance of structured table representations for tabular reasoning. The cross-modal alignment module contributes consistently across all tasks, with its removal resulting in an average performance drop of 6.4 percentage points. The numerical reasoning loss component is particularly important for TAT-QA (8.8 percentage point drop), confirming its role in enhancing precise numerical computation capabilities. The model's average performance of 71.5% demonstrates that the multimodal integration framework effectively leverages complementary information from all three modalities [8].

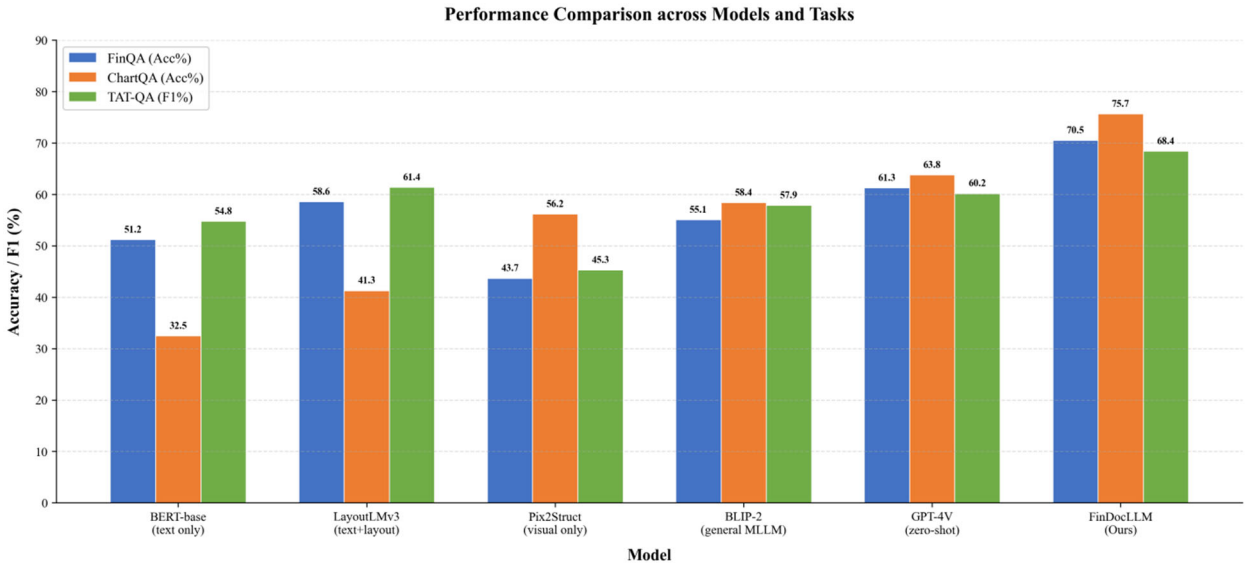

Figure 2. Performance Comparison across Models

This study presents FinDocLLM, a multimodal large language model framework designed for comprehensive financial document understanding. By integrating visual chart encoding, structured table representation, and textual narrative analysis within a unified architecture, the proposed framework achieves substantial performance improvements over existing unimodal and general-purpose multimodal baselines across three representative financial analysis benchmarks. The empirical results demonstrate that multimodal approaches consistently outperform text-only methods, confirming the critical importance of incorporating visual and tabular information in financial document analysis [5].

Our ablation analysis further reveals the complementary nature of different modality components. Visual encoding is indispensable for chart understanding, table encoding is essential for structured data reasoning, and the cross-modal alignment module serves as the bridge that enables effective information integration across modalities. The numerical reasoning loss component enhances the model's capacity for precise financial calculations, addressing a key limitation of general-purpose LLMs in the financial domain [6].

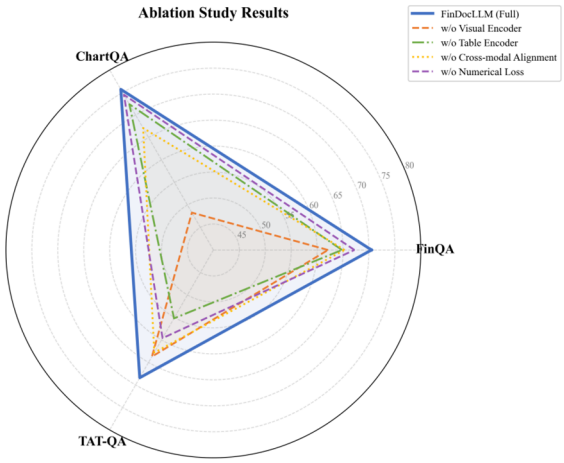

Figure 3. Ablation Study Results Visualization

However, this study acknowledges certain limitations. First, the dataset is constructed from English-language financial documents of U.S. listed companies, and the generalizability to other languages and regulatory environments remains to be verified. Second, the current framework processes documents at the page level, which may miss cross-page dependencies in longer documents. Third, the evaluation focuses on factual question answering tasks, while higher-level financial reasoning tasks such as trend forecasting and risk assessment are not covered.

Looking forward, several promising research directions emerge from this work. Future investigations would benefit from expanding the dataset to include multilingual financial documents, particularly those from Chinese capital markets. The development of more sophisticated long-document processing mechanisms would enable analysis of complete annual reports rather than individual pages. Additionally, integrating financial knowledge graphs with the multimodal LLM framework could further enhance domain-specific reasoning capabilities. Through these advancements, subsequent research can develop more comprehensive financial document understanding systems that support real-world investment analysis and regulatory compliance workflows.

6. Conclusion

This study proposes and validates FinDocLLM, a multimodal large language model framework that jointly processes textual narratives, visual charts, and structured tables for financial document understanding. Experimental results on three benchmark tasks confirm that the multimodal integration strategy consistently outperforms unimodal baselines, achieving an average accuracy of 71.5% across FinQA, ChartQA, and TAT-QA. Ablation experiments further demonstrate that each modality component contributes complementary information, with visual encoding proving most critical for chart interpretation and table encoding essential for

numerical reasoning. Nevertheless, the current study is limited to English-language documents from U.S. listed companies, and the page-level processing design may overlook cross-page dependencies in longer reports. Future work will extend the framework to multilingual financial corpora, incorporate long-document modeling mechanisms, and integrate financial knowledge graphs to strengthen domain-specific reasoning and reduce numerical hallucination in model outputs.

References

1. Chen Z, Chen W, Smiley C, et al. Finqa: A dataset of numerical reasoning over financial data[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 3697-3711.
2. Huang Y, Lv T, Cui L, et al. Layoutlmv3: Pre-training for document ai with unified text and image masking[C]//Proceedings of the 30th ACM international conference on multimedia. 2022: 4083-4091.
3. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
4. Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models [C] //International conference on machine learning. PMLR, 2023: 19730-19742.
5. Masry A, Do X L, Tan J Q, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning[C]//Findings of the association for computational linguistics: ACL 2022. 2022: 2263-2279.
6. Zhu F, Lei W, Huang Y, et al. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance[C]//Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). 2021: 3277-3287.
7. Liu F, Piccinno F, Krichene S, et al. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 12756-12770.
8. Lee K, Joshi M, Turc I R, et al. Pix2struct: Screenshot parsing as pretraining for visual language understanding[C]//International Conference on Machine Learning. PMLR, 2023: 18893-18912.