EviFiVQA: A Benchmark for Evidence-Grounded Multi-Hop Reasoning in Financial VQA

Sachin Raja[0009−0004−0385−7437], Ajoy Mondal[0000−0002−4808−8860], and C V Jawahar[0000−0001−6767−7057]

CVIT, IIIT Hyderabad, India
sachin.raja@research.iiit.ac.in, ajoy.mondal@iiit.ac.in, jawahar@iiit.ac.in

Abstract. Financial statements are critical for corporate analysis, fraud detection, and regulatory compliance, yet their complex tabular layouts, multi-year data structures, and hierarchical relationships pose significant challenges for automated reasoning. Existing Visual Question Answering (VQA) models struggle to deliver audit-ready responses due to their inability to explicitly localize evidence and reason over structured financial data. To bridge this gap, we introduce EviFiVQA, a large-scale benchmark dataset for evidence-based financial VQA, built on real-world financial statements. Leveraging FinTabNet, our dataset contains over 1.5 million question-answer pairs spanning five reasoning categories, each meticulously annotated with bounding boxes for relevant table cells to ensure traceable and interpretable responses. EviFiVQA introduces new challenges for Large Language Models (LLMs) and Vision-Language Models (VLMs) by requiring multi-step numerical reasoning, explicit evidence localization, and hierarchical financial aggregation — all within complex, irregular table layouts. We benchmark state-of-the-art models on EviFiVQA, revealing their limitations in evidence-grounded understanding of financial statements. We release EviFiVQA to drive progress in audit-ready financial AI, enhancing transparency and reliability in AI-driven financial analytics. The dataset and code are available at https://github.com/sachinraja13/EviFiVQA.

Keywords: EviFiVQA · Financial VQA · Evidence-Localization · Auditable AI · Transparent AI · Reliable AI.

1 Introduction

Financial statements serve as the primary foundation for evaluating corporate performance, detecting fraud, and meeting regulatory mandates [1,2,3]. Consequently, analysts, auditors, and regulators rely heavily on these documents to form high-stakes judgments. In such contexts, evidence traceability becomes indispensable: any conclusion drawn from a financial report must be backed by explicit references to the original tables. Without this level of evidence localization, AI-driven insights risk being opaque and can compromise the reliability of the decision-making process.

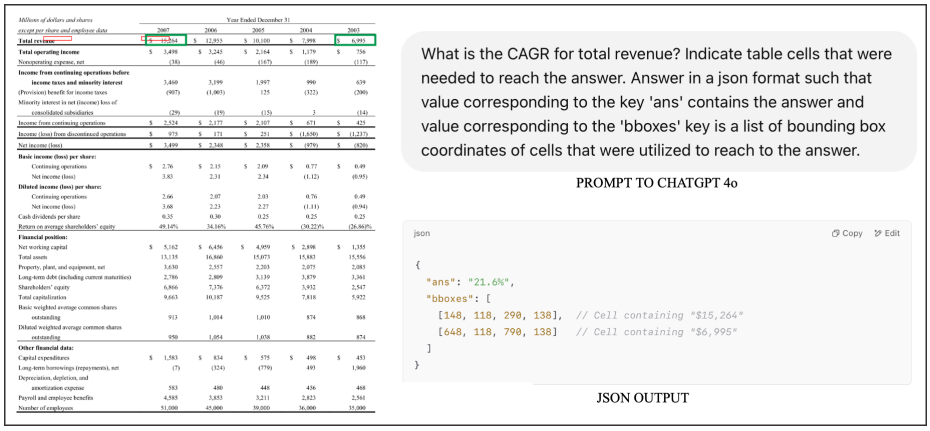
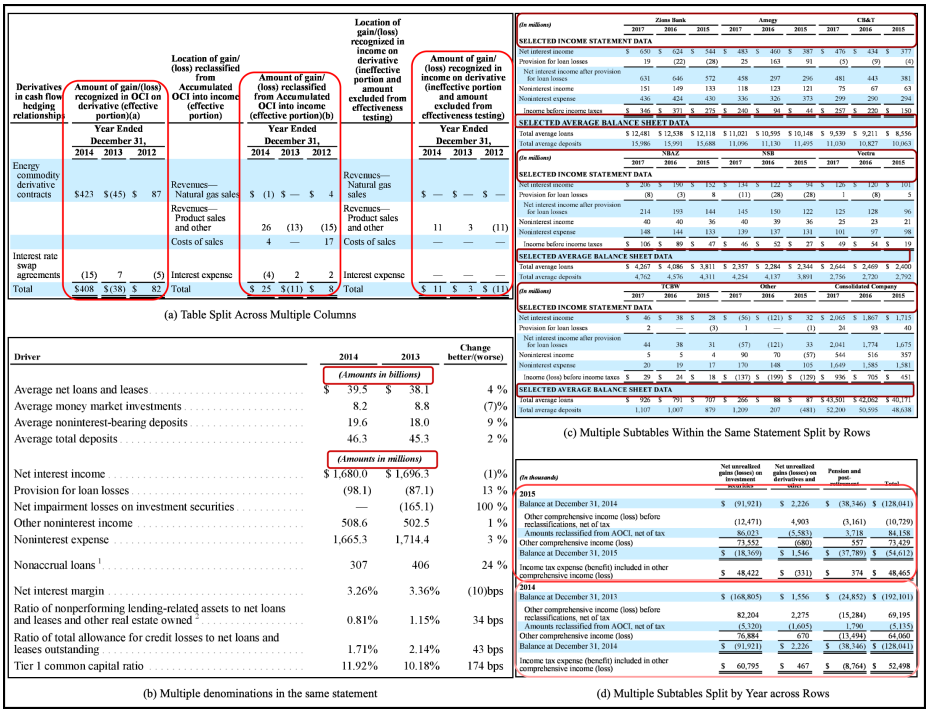


Fig. 1: Qualitative example from ChatGPT-4o model. While the answer to the question is computed correctly, the evidence as bounding boxes were off by a significant margin (green: ground-truth, red: model output).

Due to the heterogeneous and often complex nature of financial statements, which commonly feature multi-year column layouts, merged cells, footnotes, and disclaimers, traditional Optical Character Recognition (OCR) pipelines may introduce digit misreads and break the inherent link between the numeric content and the document's visual-spatial structure [4,5]. These issues create difficulties in validating which cells support a given conclusion and complicate subsequent audits. As an alternative, large Vision-Language Models [6,7,8,9,10](VLMs) can directly analyze document images, reducing the vulnerability of text-based pipelines to OCR errors and maintaining a clearer visual-spatial reference. By combining layout and textual analysis, these models can reconcile data inconsistencies in annual reports, detect misrepresentations in profit-and-loss statements, and highlight disclaimers that affect numeric interpretations. When paired with external accounting knowledge, they also identify anomalies in key financial metrics, providing a robust foundation for comprehensive risk assessments.

Importantly, by generating bounding-box evidence on the source documents, these models can facilitate transparency and auditability and interpretability. Finance professionals or regulatory bodies can directly inspect the exact segments of financial statements that inform each inference a critical feature for downstream applications such as portfolio stress testing, fraud detection, credit risk modeling, and automated compliance checks. As the scope and format of financial statements continue to evolve, large Vision-Language Models, with their capacity to unify image-level reasoning and domain-specific analysis, promise a scalable and robust approach to future challenges in financial reporting.

However, because most large-scale VLMs are trained on general-purpose datasets, their specialized performance in financial document analysis remains uncertain. Addressing this gap necessitates a structured benchmark to systematically evaluate and refine VLMs for financial tasks. To this end, we introduce



(a) Table Split Across Multiple Columns

(b) Multiple denominations in the same statement

(c) Multiple Subtables Within the Same Statement Split by Rows

(d) Multiple Subtables Split by Year across Rows

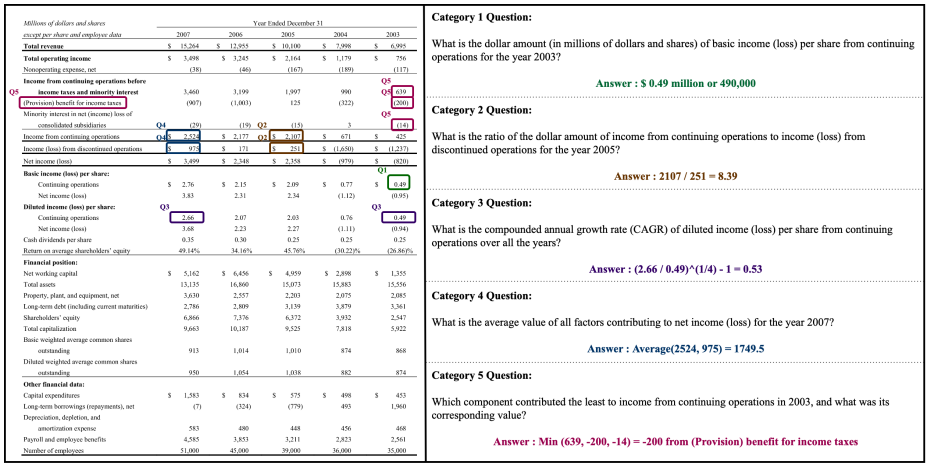
Fig. 2: Examples of complex layouts include horizontally and vertically split subtables and multiple denominations within the same statement from EviFiVQA built on top of FinTabNet [11] dataset.

EviFiVQA, a large-scale Evidence-Based Financial Visual Question Answering dataset built on top of the established FinTabNet benchmark [11]. Our dataset is designed to stimulate the development of models capable of delivering audit-ready responses rooted in explicit references to financial tables, pushing the boundaries of reliability and transparency in AI-powered financial analytics.

2 Related Work

2.1 Datasets

The emergence of Table-VQA as a research area was driven by datasets that challenged models to reason over structured tabular layouts. Early benchmarks such as WikiTableQuestions [12] and Natural Questions [13] focused on querying semi-structured tables within web pages, emphasizing compositional reasoning over cell values. While these datasets introduced fundamental concepts like cell alignment, filtering, and multi-step logical reasoning, they did not consider scanned or image-based tables, leaving a gap for subsequent research. A crucial step toward image-based Table QA involved the development of table extraction bench-



Millions of dollars and shares except per share and employee data	Year Ended December 31 2007	2006	2005	2004	2003
Total revenue	$ 15,264	$ 12,955	$ 10,100	$ 7,998	$ 6,995
Total operating income	$ 3,498	$ 3,245	$ 2,164	$ 1,179	$ 756
Nonoperating expense, net	(38)	(46)	(167)	(189)	(117)
Income from continuing operations before income taxes and minority interest	3,460	3,199	1,997	990	639
(Provision) benefit for income taxes	(907)	(1,003)	125	(322)	(200)
Minority interest in net (income) loss of consolidated subsidiaries	(29)	(19)	(15)	3	(14)
Income from continuing operations	$ 2,524	$ 2,177	$ 2,107	$ 671	$ 425
Income (loss) from discontinued operations	$ 975	$ 171	$ 251	$ (1,650)	$ (1,237)
Net income (loss)	$ 3,499	$ 2,348	$ 2,358	$ (979)	$ (820)
Basic income (loss) per share:					
Continuing operations	$ 2.76	$ 2.15	$ 2.09	$ 0.77	$ 0.45
Net income (loss)	3.83	2.31	2.34	(1.12)	(0.95)
Diluted income (loss) per share:					
Continuing operations	2.66	2.07	2.03	0.76	0.49
Net income (loss)	3.68	2.23	2.27	(1.11)	(0.94)
Cash dividends per share	0.35	0.30	0.25	0.25	0.25
Return on average shareholders' equity	49.14%	34.16%	45.76%	(30.22)%	(26.86)%
Financial position:					
Net working capital	$ 5,162	$ 6,456	$ 4,959	$ 2,898	$ 1,355
Total assets	13,135	16,860	15,073	15,883	15,556
Property, plant, and equipment, net	3,630	2,557	2,203	2,075	2,085
Long-term debt (including current maturities)	2,786	2,809	3,139	3,879	3,361
Shareholders' equity	6,866	7,376	6,372	3,932	2,547
Total capitalization	9,663	10,187	9,525	7,818	5,922
Basic weighted average common shares outstanding	913	1,014	1,010	874	868
Diluted weighted average common shares outstanding	950	1,054	1,038	882	874
Other financial data:					
Capital expenditures	$ 1,583	$ 834	$ 575	$ 498	$ 453
Long-term borrowings (repayments), net	(7)	(324)	(779)	493	1,960
Depreciation, depletion, and amortization expense	583	480	448	456	468
Payroll and employee benefits	4,585	3,853	3,211	2,823	2,561
Number of employees	51,000	45,000	39,000	36,000	35,000

Category 1 Question:

What is the dollar amount (in millions of dollars and shares) of basic income (loss) per share from continuing operations for the year 2003?

Answer : $ 0.49 million or 490,000

Category 2 Question:

What is the ratio of the dollar amount of income from continuing operations to income (loss) from discontinued operations for the year 2005?

Answer : 2107 / 251 = 8.39

Category 3 Question:

What is the compounded annual growth rate (CAGR) of diluted income (loss) per share from continuing operations over all the years?

Answer : (2.66 / 0.49)^(1/4) - 1 = 0.53

Category 4 Question:

What is the average value of all factors contributing to net income (loss) for the year 2007?

Answer : Average(2524, 975) = 1749.5

Category 5 Question:

Which component contributed the least to income from continuing operations in 2003, and what was its corresponding value?

Answer : Min (639, -200, -14) = -200 from (Provision) benefit for income taxes

Fig. 3: Examples of different categories of question from a sample image.

marks. PubTabNet [14] and the ICDAR 2019 Table Recognition Dataset have been instrumental in evaluating table detection and structure recognition, facilitating progress in parsing scanned documents. TableVQA-Bench [15] extends these efforts by aggregating table-image sources from diverse domains, such as receipts, scientific reports, and business documents, making it a strong testbed for model generalization. Datasets targeting reasoning over multimodal tabular data have further propelled the field. TAT-QA [16] integrates both tabular and textual evidence for financial question answering, emphasizing hybrid reasoning. FeTaQA [17] encourages free-form answer generation by requiring models to synthesize information across multiple cells. More recently, TableBench [18] and VQAonBD [19] have introduced multi-hop reasoning over multiple tables, while TQA-Bench [20] evaluates LLMs on symbolic joins and large-scale retrieval. Another significant dataset is FinQA [21], which focuses on numerical reasoning over financial reports. It combines natural language questions with both textual and tabular evidence, requiring models to perform arithmetic computations, aggregation, and multi-step reasoning. FinQA highlights the complexities involved in financial document question answering, where precise numeric reasoning and contextual understanding are crucial. These datasets highlight the evolving complexity of Table-VQA tasks and the need for robust models capable of reasoning over heterogeneous tabular data.

2.2 Methods

Methods for Table-VQA have evolved from semantic parsing to neural network-based approaches, incorporating both text-based and image-centric techniques. Early methods such as TAPAS [22] adapted BERT-style transformers for structured tables by encoding row and column embeddings, enabling models to perform aggregation operations directly on text-based tables. Further refinements,

	Num Images	Category 1	Category 2	Category 3	Category 4	Category 5	Total
Training Set	34,790	481,025	285,881	200,482	117,138	165,471	1,249,997
Test Set	3,654	49,386	30,127	21,936	11,678	15,137	128,264
Validation Set	3,943	53,352	31,590	22,583	12,566	16,983	137,074
Total	42,387	583,763	347,598	245,001	141,382	197,591	1,515,335

Table 1: Distribution of questions across varying difficulty categories.

such as intermediate pretraining on counterfactual data [23], enhanced model robustness. However, these approaches assumed structured, digital input and struggled with visual tables containing irregular layouts or OCR errors. With the rise of image-centric Table-VQA, table detection and OCR pipelines became integral. TATR [24] demonstrated how transformer-based architectures could handle OCR-extracted text, while CascadeTabNet [25] improved structure recognition via CNN and RNN cascades. TabIQA [26] addressed challenges in business document QA by integrating region-based CNNs with contextual question answering models. Similarly, Xue *et al.* [27] designed a system that first extracts structured text before applying neural question answering, highlighting the difficulty of reconciling OCR errors with reasoning tasks. Recent advances leverage multimodal fusion to enhance Table-VQA. M-TBQA [28] combines textual embeddings with layout-derived visual cues such as bounding boxes and font sizes. Graph-based models like Zhang *et al.* [29] explicitly model row-column relationships, improving interpretability. Hybrid methods such as HYTREL [30] use hypergraph representations to capture complex table interactions, ensuring robustness to structural variations. The advent of large language models (LLMs) has introduced new paradigms in Table-VQA. TAPEX [31] pre-trains models on synthetic SQL queries to simulate symbolic reasoning. ReAcTable [32] incorporates chain-of-thought prompting, guiding models to sequentially query, update, and verify answers. Multi-agent strategies, such as AutoTQA [33], decompose complex queries into subtasks handled by specialized LLMs, improving answer interpretability. Furthermore, TableLlama [34] and OmniTab [35] push generalist models toward handling diverse table formats, including scanned documents, structured databases, and hybrid tabular-textual sources.

Despite these advances, Table-VQA models still face challenges with noisy OCR outputs, multi-hop reasoning across tables, and domain-specific adaptation. Ongoing work explores solutions such as denoising retrieval [36], hierarchical transformers for structured document parsing [37], and domain-specific benchmarks like FREB-TQA [38]. Future research must bridge the gap between structured parsing and multimodal reasoning, ensuring robust, scalable, and interpretable Table-VQA systems for real-world applications.

3 EviFiVQA Dataset

We construct EviFiVQA from publicly available S&P 500 earnings reports (1999-2019), leveraging FinTabNet [11]. These multi-page PDFs mix textual discussion with tabular data, and FinTabNet's table annotations allow us to automate

Algorithm 1 Dataset Annotation Pipeline

Input: Table image or document
Output: Annotated QA pairs with bounding boxes
Step 1: Datatype Classification
  Classify cells into categories: year, date, numerical, percentage, empty, text.
Step 2: Structural Analysis
  Identify data rows, column headers, and row headers using heuristic rules.
  Detect nested sub-tables and segment accordingly
Identify Column Headers
  Identify rows suspected to contain column headers and isolate the cells.
  Determine the start data column by skipping row/column header cells.
  Parse cell content to detect specific patterns (e.g., year, denomination, numeric).
  Classify each row: spans multiple columns or is purely numeric/textual.
  If no headers are identified, assign generic labels.
Identify Document Tree Structure
  Find partial sums in each column where consecutive entries form the next entry.
  Remove groupings that are subsets of larger ones.
  Combine results across columns, reinforcing on textual cues (headers/footers).
  Normalize scores and prune groupings below a certain threshold.
  Generate the tree structure recursively representing table hierarchy.
Identify Row Headers
  Extract indentation and numeric groupings.
  Group adjacent rows with similar indentation.
  Relate row text to numeric sum groupings.
  Detect sub-headers or net totals.
  Consolidate repeated headings under single reference.
  Identify group headers and associate them with respective row headers.
  Merge sub-headers with nearest total rows.
  Assign final row headers or placeholders ensuring no duplication.
Step 4: QA Generation
  Generate template-based question-answer pairs or each category o questions.
  Paraphrase questions using LLaMA 3.2 7B to ensure linguistic coherence.
Return: Final structured dataset with annotated QA pairs.

dataset creation for evidence-based visual question answering (EVQA), validated through manual checks. Each QA pair specifies all table cells involved in the answer and includes bounding boxes as evidence.

The FinTabNet tables pose visual challenges of high row/column density, nested structures, and multi span cells (Figure 2). Additionally, core financial statements (balance sheets, income statements, cash flow statements) follow a tree-like hierarchy: broad categories (e.g., assets, liabilities) are roots, subcategories are intermediates, and line items are leaves. This structure facilitates aggregation, drill-down analysis, and dependency tracking. Consequently, Evi-FiVQA covers a spectrum of QA tasks, from simple data extraction to more complex queries requiring a semantic grasp of the table's hierarchical layout.

Dataset Annotation: Our dataset annotation framework adopts a bottom-up methodology, commencing with a functional analysis of the table, followed by the extraction of its hierarchical structure, and culminating in the generation of question-answer (QA) pairs leveraging the extracted metadata. The dataset

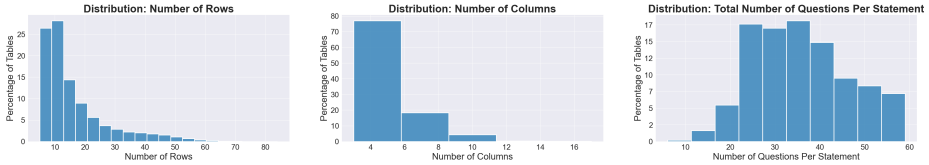


Fig. 4: Distribution of image-level parameters. Left: number of rows; Middle: number of columns; Right: number of questions per statement.

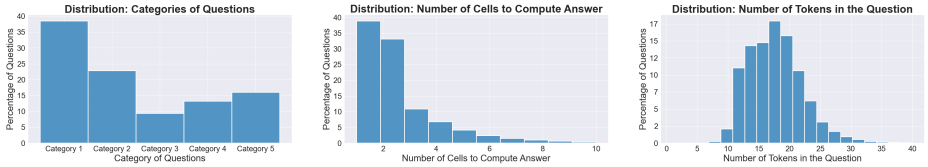


Fig. 5: Distribution of question-level parameters. Left: category of questions; Middle: number of cells required to compute the answer; Right: number of tokens in the questions

annotation process follows a structured pipeline to ensure a comprehensive and unambiguous representation of tabular data as shown in Algorithm 1.

We categorize EviFiVQA's questions into five types based on the complexity of reasoning:

1. **Extractive (Category 1):** Direct lookups (e.g., *"Based on this financial report, what is the net revenue of the organization in 2008?"*). The answer is found by intersecting the correct row and column.
2. **Ratio-Based (Category 2):** Requires division across rows or columns (e.g., *"What is the ratio of the net revenue in 2007 to 2008 for this organization?"*).
3. **All-Years Aggregation (Category 3):** Demands operations like *avg*, *min*, *max*, or *compounded growth* over a row across multiple years; (e.g., *"For this organization, what is the compounded annual growth rate across all the years?"*).
4. **Hierarchical Aggregation (Category 4):** Involves summing or averaging hierarchical sub-entries (e.g., *"What is the average value of all the factors contributing to net income for the year 2008?"*).
5. **Key-Value Extraction (Category 5):** Identifies the *key* (textual row header) and its *value*, typically within a parent-child financial structure (e.g., *"Which factor contributed most towards net income 2008, and what is its corresponding value?"*).

Each QA pair is linked to bounding boxes along with the individual values for every cell needed in computing or justifying the answer. Figure 3 shows examples of these question types for a sample image. Refer Table 1 and Figures 4 and 5 for detailed statistical understanding of EviFiVQA dataset.

Validating the Dataset: We manually validated our dataset by randomly sampling between 2 and 5 questions from each category for every image in the

collection. After selecting these questions, we compared the annotated answers and bounding boxes to the original documents, with subject matter experts (SMEs) — experienced in finance or accounting-verifying whether each question was answered correctly and whether the corresponding row and column headers were accurately mapped. We found that there was 97.37% agreement between the SME evaluations and our existing annotations, indicating a high level of accuracy in the dataset. In cases where discrepancies arose, we documented and removed the entire financial statement from the dataset, further refining the annotations and strengthening our confidence. Out of the 2.63% erroneous cases, 1.2% were due to ambiguity induced by paraphrasing the question, 0.4% cases had ambiguity in answers because of multi-level numerical hierarchy and remaining 1% cases were due to boundary cases not properly handled by the automated question generation process.

In parallel, we implemented automated checks to assess hierarchical consistency and arithmetic correctness at scale. We verified that bounding boxes aligned with the original FinTabNet [11] annotations, and that row and column headers matched their respective data points. We also recalculated sums, averages, ratios, and growth rates to confirm the correctness of the annotated values, flagging any discrepancies that exceeded our predefined tolerance. Additionally, we employed semantic similarity using SentenceBERT [39] to evaluate the paraphrased questions on a similarity threshold of 0.85, ensuring they faithfully retained the meaning of the original queries. Whenever we detected mismatches — whether in bounding box alignments, arithmetic calculations, or textual annotations — we removed those annotations. for further manual inspection and resolved them accordingly. This dual approach, combining both human expertise and automated validation, helped ensure the dataset's overall robustness and reliability.

Novelty of EviFiVQA Existing financial visual question answering (VQA) datasets, such as FinQA [21] and FinTabNetQA [15], have significantly advanced the field by introducing complex numerical reasoning and table-based question-answering tasks. FinQA provides 8.5K expert-annotated QA pairs over financial reports, focusing on numerical reasoning with structured and unstructured evidence. However, it primarily emphasizes textual and tabular data without incorporating visual elements or detailed hierarchical structures. FinTabNetQA, derived from the FinTabNet dataset, extends table-based QA by generating questions using large language models and rendering tables into images. While it addresses visual aspects, it lacks manual validation and does not explicitly model the underlying tree structures of financial tables. In contrast, our EviFiVQA dataset uniquely integrates the inherent hierarchical tree structures of financial documents to curate aggregation and key-value extraction questions (refer Table 2). This approach enables models to perform complex multi-hop reasoning by understanding relationships between financial categories and subcategories, a feature not present in previous datasets. Additionally, EviFiVQA provides comprehensive visual elements and manual validation, with each QA pair meticulously annotated with bounding boxes for all relevant table cells, of-

Table 2: Comparison of Table QA Datasets

Dataset (Year)	Size	Source	Annotation	Unique Points
WikiTable-Questions (2015)	2K tables; 22K QA	Wikipedia (HTML)	Crowdsourced (Natural Qs)	Compositional, multi-step
WikiSQL (2017)	120K (tables + SQL)	Wikipedia	Crowdsourced (NL to SQL)	SQL generation, conditions + aggregations
TabFact (2020)	16.5K	Wikipedia	Auto-gen + human check	Fact verification (True/False), logical reasoning
HybridQA (2020)	70K QA	Tables and text	Crowdsourced (multi-hop)	Combines table + multi-hop reasoning
TAT-QA (2020)	16K QA	10-K financials	Experts and Crowdsourced	Numeric multi-step, real-world finance
FeTaQA (2021)	10.3K QA	Wikipedia (semi-structured)	Crowdsourced (free-form)	Longer answers, multi-entity aggregation
FinQA (2021)	8K QA	Earnings reports and FinTabNet	Finance experts	Multi-step numerical calculations
AIT-QA (2022)	10-K forms (5 airlines)	SEC EDGAR	Human (airline KPIs)	Airline-specific tables, KPI queries
TableBench (2023)	886 + 19.7K QA	Real-world tables	ChatGPT and Human checked	Numeric + multi-domain coverage
***EviFiVQA** (2025)*	***1.5M QA (5 categories) + Evidence***	***FinTabNet***	***Automated with Expert Validation***	***Semantic, multi-step complex aggregations***

fering explicit evidence for the answers. This design enhances interpretability and addresses real-world challenges such as high-density tables, multi-span cells, and nested structures, positioning EviFiVQA as a novel and robust benchmark for evidence-based visual question answering in the financial domain.

4 Evaluation

We propose a new evaluation protocol for evidence-based VQA on financial tables, designed to capture the correct identification of all relevant cells (their bounding boxes) and the correctness of the final answer. In particular, we split the overall metric into two main components: an *Evidence Score*, which measures how accurately the system detects the necessary table cells, and an *Answer Score*, which assesses the final predicted answer without over-penalizing minor OCR mistakes. We then combine these two parts into a single, interpretable metric.

4.1 Evidence Score

Answers to many questions in financial documents require gathering evidence from multiple table cells, each with its bounding box. To handle this scenario, we model the set of relevant bounding boxes. Let $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$ be the set of

ground-truth bounding boxes required for the answer, and let $\mathcal{P} = \{p_1, \ldots, p_{|\mathcal{P}|}\}$ be the set of predicted boxes. At a set level, we treat the matching problem between $\mathcal{G}$ and $\mathcal{P}$.

Intersection-over-Union (IoU): For any predicted box p and ground-truth box g, we define:

$$\text{IoU}(p, g) \;=\; \frac{|p \cap g|}{|p \cup g|},$$

where $|p \cap g|$ is the area of overlap and $|p \cup g|$ is the area of the union of the two boxes.

Set Matching: We perform a *one-to-one* matching based on IoU to identify correct predicted boxes.

1. For each predicted box p, find its best-matching ground-truth box g^* by maximizing $\text{IoU}(p, g)$.
2. If $\max_{g \in \mathcal{G}} \text{IoU}(p, g) \geq \delta$, we count this as a true positive (TP), and we mark g^* as matched so it cannot be used for another predicted box.
3. Boxes in $\mathcal{P}$ that do not satisfy IoU $\geq \delta$ with any unmatched ground-truth box are considered false positives (FP).
4. Boxes in $\mathcal{G}$ that remain unmatched after processing all $\mathcal{P}$ are considered false negatives (FN).

Evidence Precision (EP) & Evidence Recall (ER): With the above matching mechanism:

$$\text{EP} \;=\; \frac{\text{TP}}{|\mathcal{P}|}, \quad \text{ER} \;=\; \frac{\text{TP}}{|\mathcal{G}|}.$$

Intuitively, EP is low if the model predicts many superfluous boxes, while ER is low if it misses too many required boxes.

Evidence F1: We combine EP and ER into an F1 measure:

$$\text{EvidenceF1} \;=\; 2 \;\times\; \frac{\text{EP} \times \text{ER}}{\text{EP} + \text{ER}}.$$

Because some questions require aggregating multiple table cells, this *EvidenceF1* metric rewards precise yet comprehensive identification of all necessary bounding boxes. We use a modest IoU threshold δ (0.5) to allow for small misalignments in bounding-box predictions without harshly penalizing near-correct localizations.

4.2 Answer Score

During our evaluation, the model receives only the document image and a question as inputs, and it must produce an answer that we compare to the ground truth. We aggregate the scores over the entire evaluation dataset to measure performance. Below, we detail two main components of our scoring protocol: *averaged absolute deviation for numeric answers*, and *averaged normalized Levenshtein similarity* (ANLS) [40,41].

Averaged Absolute Deviation for Numeric Answers: To account for exact numerical matches or slight numeric deviations, we measure how far a predicted value is from the ground truth. Let x be the correct numeric value and $\widehat{x}$ be the model's prediction. We assign a score of 0 if $|\widehat{x} - x|$ exceeds $|x|$ (i.e., more than 100% error). Otherwise, we define:

$$\text{DeviationScore} \;=\; 1 \;-\; \frac{\big|\widehat{x} - x\big|}{\big|x\big|}. \tag{1}$$

We then average this score over all questions requiring a numeric answer. This yields a softer evaluation than simple accuracy, especially for minor numerical differences.

Averaged Normalized Levenshtein Similarity (ANLS): Since our model relies on OCR to extract tokens from the document image, we also incorporate a measure that softly handles minor recognition errors. We employ the *averaged normalized Levenshtein similarity* (ANLS) [40,41]. Suppose each question q_i has one or more ground-truth answers $\{a_{ij}\}$, and let $\widehat{o}_{q_i}$ be the predicted answer. Denoting by $NL(\cdot,\cdot)$ the normalized Levenshtein distance between two strings (ranging from 0 to 1), ANLS is computed as:

$$\begin{aligned} \text{ANLS} \;&=\; \frac{1}{N}\sum_{i=1}^{N}\Big(\max_{j}\; s\big(a_{ij}, \widehat{o}_{q_i}\big)\Big), \\ s\big(a_{ij}, \widehat{o}_{q_i}\big) \;&=\; \begin{cases} 1 - NL\big(a_{ij}, \widehat{o}_{q_i}\big), & \text{if } NL\big(a_{ij}, \widehat{o}_{q_i}\big) < \tau, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{2}$$

where τ (0.5) is a threshold to discount large edit distances. If $NL(a_{ij}, \widehat{o}_{q_i}) \geq \tau$, we consider the error significant enough to be an incorrect prediction rather than a minor OCR slip.

For numeric answers, we combine the Deviation Score in Eq. (1) and ANLS in Eq. (2) by taking their L2 norm. Both measures lie in $[0,\ 1]$, so their L2 norm also lies in $[0,\ \sqrt{2}]$ but we normalize back to $[0,\ 1]$ if desired:

$$\text{AnswerScore}_{\text{numeric}} \;=\; \frac{\sqrt{\big(\text{Deviation Score}\big)^2 \;+\; \big(\text{ANLS}\big)^2}}{\sqrt{2}}.$$

For textual answers, we rely solely on the ANLS value as the final measure. By leveraging exact accuracy, numeric deviation, and fuzzy matching, our approach remains robust to OCR noise while still rewarding precise answers for business-document VQA.

Final Answer Score: If an answer is purely textual, we adopt the fuzzy matching score. If it is numeric, we compute numScore directly. For answers that combine textual and numeric content, we can use a simple weighted average or product of their respective scores, depending on the nature of the task. We denote the resulting overall measure by AnswerScore.

4.3 Combined Metric

Our final metric reflects correct cell identification and the correctness of the produced answer. We propose combining the two components as a product:

$$\text{TotalScore} = \text{EvidenceF1} \times \text{AnswerScore}. \tag{3}$$

In this formulation, a perfect bounding-box detection (EvidenceF1 = 1) and a perfect answer (AnswerScore = 1) yields a maximum TotalScore of 1.0. A model that identifies only a subset of the relevant cells or provides an approximate final answer is scored accordingly (i.e., one of the factors diminishes, thereby reducing the product). This multiplicative scheme effectively balances precision and recall in the detection stage with the final correctness of the answer.

5 Experiments

5.1 Experiment Setup

We use large vision-language models as baselines to benchmark the proposed dataset. We conduct experiments on a single M1-Ultra Chip with 64 GB unified memory using the framework of Apple's MLX [42]. We conduct our experimental study in two setups — (i) *Zero-shot Setting:* we use a pre-trained model to directly predict the output through prompt engineering, and (ii) *Fine-tuned Setting:* we fine-tune the pre-trained model on one randomly selected QA pair from each category of questions for each document image in training set for two epochs because of resource limitations to identify the best-performing baseline. We then fine-tune this baseline on the entire training dataset for one epoch. For finetuning the model, we use the evidence bounding boxes and the individual answer annotations to generate chain of reasoning tokens as a part of the ground-truth. We use a learning rate of 5e-5 with a cosine decay learning schedule and 500 steps to warm up the learning rate for all our fine-tuning experiments. Further, we fine-tune the base model, keeping the vision backbone frozen and using 4-bit quantized LoRA for all the projection matrices.

Method	**#Params**	**Deviation Score**	**ANLS Score**	**Answer Score**	**Evidence Score**	**Combined Score**
QWEN2-VL [6]	7B	0.28	0.35	0.32	0.06	0.02
LLaVA [7]	7B	0.25	0.33	0.29	0.04	0.01
Pixtral [8]	12B	0.29	0.36	0.33	0.07	0.02
Llama-3.2-vision [10]	11B	0.30	0.38	0.34	0.07	0.02
DeepSeek-VL [9]	7B	0.33	0.47	0.41	0.09	0.04

Table 3: Performance comparison of vision-language models in a zero-shot setting. All scores are averaged for each QA pair across all categories of questions. *Combined Score* multiplies *Answer Score* and *Evidence Score*, reflecting the requirement for both a correct answer and correct evidence.

We experiment with vision-language models implemented as a part of MLX-VLM [43] library and support fine-tuning. We compare QWEN2-VL [6], LLaVA [7], Pixtral [8], DeepSeek-VL [9] and Llama-3.2-vision [10]. For each of these models, we use the same prompt — "Question String + Answer in a JSON format such that value corresponding to the key 'ans' contains the answer and value corresponding to the 'bboxes' key is a list of bounding box coordinates of the cells that were used to compute the answer".

5.2 Experimental Results

In Table 3, the zero-shot performance of these vision-language models remains uniformly low across all metrics, emphasizing the difficulty of extracting precise information from complex financial statements with no domain-specific adaptation. Even the best-performing model (DeepSeek-VL) achieves only 0.33 numerical deviation and 0.09 Evidence Score, resulting in a Combined Score of 0.04. Such low accuracy underscores two central challenges. First, identifying the correct numeric or textual answer in highly dense layouts requires specialized reasoning about rows, columns, and multi-year structures that generic pre-training does not provide. Second, the models typically fail to highlight the relevant cells, as indicated by Evidence Scores below 0.10 for all methods. This issue is particularly problematic in a financial context, where each value should be explicitly tied to its source for audit ability and regulatory compliance. The net effect is that zero-shot VLMs struggle to achieve reliable performance, as Combined Scores barely exceed 0.01 for most baselines.

Method	**#Params**	**Deviation Score**	**ANLS Score**	**Answer Score**	**Evidence Score**	**Combined Score**
QWEN2-VL [6]	7B	0.38	0.50	0.44	0.18	0.08
LLaVA [7]	7B	0.35	0.47	0.41	0.11	0.05
Pixtral [8]	12B	0.40	0.51	0.46	0.19	0.09
Llama-3.2-vision [10]	11B	0.43	0.55	0.49	0.21	0.10
DeepSeek-VL [9]	7B	0.48	0.63	0.56	0.28	0.16

Table 4: Performance comparison of vision-language models after fine-tuning on one randomly sampled QA pair from each category for each image for two epochs.

When exposed to only a single QA pair per category in each image, every model sees a notable improvement in both exactness of answers and bounding-box localization (see Table 4). DeepSeek-VL, for instance, increases its numerical deviation score from 0.33 to 0.48 and its Evidence Score from 0.09 to 0.28, nearly quadrupling the combined Score to 0.16. We, therefore, choose DeepSeek-VL as the baseline for our detailed experiments. While these numbers remain substantially lower than required for real-world adoption, they illustrate how even limited domain-relevant data can guide the model toward better numeric reasoning

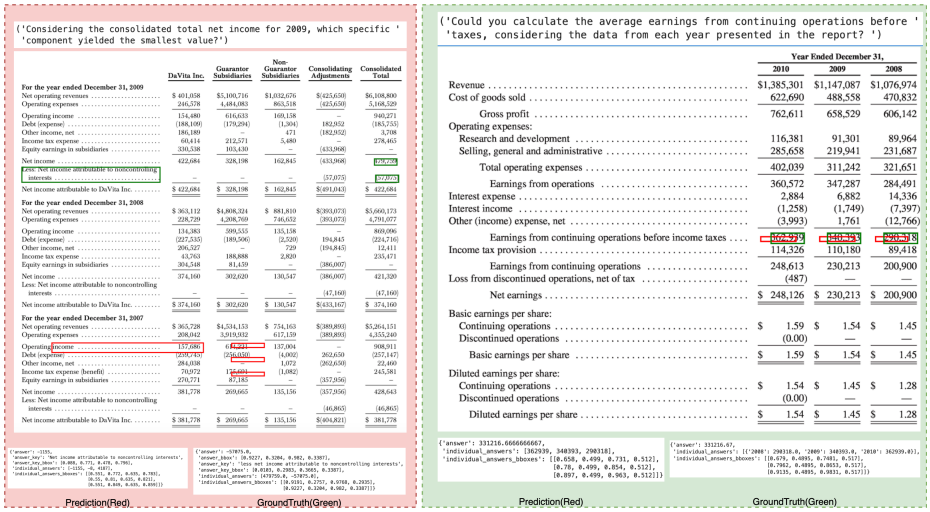


Fig. 6: Qualitative example from finetuned baseline on 2 different categories of questions from two different samples in the validation set.

and visual grounding in complex tabular layouts. Nonetheless, a numerical deviation score of 0.48 and exact match score of 0.28 still implies that roughly three out of four answers are incorrect, and an Evidence Score around 0.28 means most relevant cells are still overlooked. These findings reinforce the need for continued investigation into domain-specific representation learning, especially given the intricate hierarchical structures, multi-year columns, and specialized calculations found in financial statements. Models that can robustly handle ratio analysis, partial sums across rows, and multi-step numeric derivations are crucial for delivering accurate, audit-ready outputs, and the relatively low post fine-tuning scores confirm that bridging this gap is far from trivial.

Category of Questions	Deviation Score	ANLS Score	Answer Score	Evidence Score	Combined Score
Category 1	0.62	0.82	0.73	0.40	0.29
Category 2	0.54	0.77	0.67	0.34	0.23
Category 3	0.57	0.81	0.70	0.36	0.25
Category 4	0.50	0.72	0.62	0.29	0.18
Category 5	0.39	0.63	0.52	0.23	0.12
Averaged	0.53	0.75	0.65	0.32	0.21

Table 5: Approximate performance of DeepSeek-VL on each question category after fine-tuning on the entire training dataset for one epoch. All scores are averaged for each QA pair across each categories of questions individually.

The results in Table 5 indicate that after fine-tuning with the entire training dataset for one epoch, Category 1 (extractive questions) is significantly more

tractable than the others, with both high ANLS (0.82) and a relatively strong Evidence Score (0.40), reflecting the ease of directly mapping a single cell to the correct answer and bounding box when no multi-step operations are needed. In contrast, Category 2 and Category 3, which require ratio computations and multi-year aggregation, respectively, show moderate performance gains in ANLS and Answer Score but still dip below Category 1 in Exact-Match and bounding-box localization, likely due to the extra step of identifying multiple cells before numerical operations. Category 4, requiring hierarchical aggregation across multiple table segments, sees lower numerical deviation score (0.50), and the Evidence Score declines to 0.29, underlining the difficulty of correctly localizing every component in a more involved computation. Category 5, which combines the aggregation challenge with key-value extraction, is consistently the hardest, posting an numerical deviation of 0.39 and a Combined Score of 0.12. This drop is consistent with the model's struggle to pair the correct numeric figure with its corresponding row header in a multi-hop inference setting. Overall, these category-level results underscore how progressively higher-order reasoning-such as tracking multiple rows and columns or merging hierarchical sub-totals-places an increasing burden on the model's numeric capabilities and ability to ground answers in specific table regions. Figure 6 shows two qualitative examples, one negative and one positive for reference.

6 Conclusion

In conclusion, EviFiVQA provides a novel and rigorous platform for training and evaluating evidence-grounded reasoning in the financial domain. By focusing on multi-hop numeric computations, hierarchical table structures, and explicit bounding-box localization, our benchmark highlights the unique challenges present in financial statements-especially those requiring audit-ready, transparent answers. Our experiments with leading vision-language models reveal the considerable gap between general-purpose VQA performance and the specialized needs of real-world financial analysis. We hope that EviFiVQA spurs new research directions in interpretable modeling, domain-adaptive training, and advanced numerical reasoning, ultimately paving the way for more transparent and reliable AI systems in finance.

Acknowledgments

This work is supported by the MeitY Government of India, through the NLTM Bhashini (https://bhashini.gov.in/) project.

References

1. Chunkai Ding. Cyber Security Intelligence and Analytics, The 5th International Conference on Cyber Security Intelligence and Analytics (CSIA 2023), Volume 1. pages 160–169, 4 2023. 1

2. Rongxu Liu. Applications of New Technologies to Recognition of Financial Statement Fraud. *2020 International Conference on Computer Communication and Network Security (CCNS)*, 00:22–25, 8 2020. 1
3. Financial Accounting, Reporting & Analysis. 5 2017. 1
4. Sachin Raja, Ajoy Mondal, and C. V. Jawahar. Table structure recognition using top-down and bottom-up cues. In *ECCV*, 2020. 2
5. Hangdi Xing, Feiyu Gao, Rujiao Long, Jiajun Bu, Qi Zheng, Liangcheng Li, Cong Yao, and Zhi Yu. Lore: Logical location regression network for table structure recognition. *arXiv preprint arXiv:2303.03730*, 2023. 2
6. Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 12, 13
7. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 12, 13
8. Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 2, 12, 13
9. Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2, 12, 13
10. Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 12, 13
11. Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706, 2021. 3, 5, 8
12. Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 3
13. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 3
14. Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. *arXiv*, 2019. 4
15. Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. TableVQA-Bench: A Visual Question Answering Benchmark on Multiple Table Domains. *arXiv*, 2024. 4, 8
16. Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. *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)*, pages 3277–3287, 2021. 4
17. Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech KryÅciÅski, Hailey Schoelkopf, Riley Kong, Xiangru Tang,

Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022. 4

18. Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, and Zhoujun Li. TableBench: A Comprehensive and Complex Benchmark for Table Question Answering. *arXiv*, 2024. 4
19. Sachin Raja, Ajoy Mondal, and CV Jawahar. Icdar 2023 competition on visual question answering on business document images. In *International Conference on Document Analysis and Recognition*, pages 454–470. Springer, 2023. 4
20. Zipeng Qiu, You Peng, Guangxin He, Binhang Yuan, and Chen Wang. TQA-Bench: Evaluating LLMs for Multi-Table Question Answering with Scalable Context and Symbolic Extension. *arXiv*, 2024. 4
21. Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021. 4, 8
22. Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*, 2020. 4
23. Julian Martin Eisenschlos, Syrine Krichene, and Thomas Müller. Understanding tables with intermediate pre-training. *arXiv preprint arXiv:2010.00571*, 2020. 5
24. Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642, 2022. 5
25. Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 572–573, 2020. 5
26. Phuc Nguyen, Nam Tuan Ly, Hideaki Takeda, and Atsuhiro Takasu. TabIQA: Table Questions Answering on Business Document Images. *arXiv*, 2023. 5
27. Wenyuan Xue, Siqi Cai, Wen Wang, Qingyong Li, Baosheng Yu, Yibing Zhan, and Dacheng Tao. A Question Answering System for Unstructured Table Images. *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2783–2785, 2021. 5
28. Jingwen Zeng, Zhidong Wu, Rongrong Zheng, Wenting Xue, Chenhui Wang, Xiaoyang Yu, Tao Zhang, Shaozu Yuan, and Tiangang Zhu. M-TBQA: Multimodal Table-Based Question Answering. *2023 4th International Conference on Machine Learning and Computer Application*, pages 227–231, 2023. 5
29. Tingyi Zhang, Yubo Liu, Jie Wu, and Jianming Liao. Answering questions over tables based on TAPAS and graph attention model. *Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing*, pages 8–13, 2022. 5
30. Pei Chen, Soumajyoti Sarkar, Leonard Lausen, Balasubramaniam Srinivasan, Sheng Zha, Ruihong Huang, and George Karypis. HYTREL: Hypergraph-enhanced Tabular Data Representation Learning. *arXiv*, 2023. 5

31. Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. TAPEX: Table Pre-training via Learning a Neural SQL Executor. *arXiv*, 2021. 5
32. Yunjia Zhang, Jordan Henkel, Avrilia Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M Patel. ReAcTable: Enhancing ReAct for Table Question Answering. *Proceedings of the VLDB Endowment*, 17(8):1981–1994, 2024. 5
33. Jun-Peng Zhu, Peng Cai, Kai Xu, Li Li, Yishen Sun, Shuai Zhou, Haihuang Su, Liu Tang, and Qi Liu. AutoTQA: Towards Autonomous Tabular Question Answering through Multi-Agent Large Language Models. *Proceedings of the VLDB Endowment*, 17(12):3920–3933, 2024. 5
34. Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. TableLlama: Towards Open Large Generalist Models for Tables. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6024–6044, 2024. 5
35. Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. OmniTab: Pretraining with Natural and Synthetic Data for Few-shot Table-based Question Answering. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 932–942, 2022. 5
36. Deokhyung Kang, Baikjin Jung, Yunsu Kim, and Gary Geunbae Lee. Denoising Table-Text Retrieval for Open-Domain Question Answering. *arXiv*, 2024. 5
37. Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Avantika Sharma, Neeraj Gunda, and Sai Krishna Bala. Evaluation of Table Representations to Answer Questions from Tables in Documents : A Case Study using 3GPP Specifications. *arXiv*, 2024. 5
38. Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. FREB-TQA: A Fine-Grained Robustness Evaluation Benchmark for Table Question Answering. *arXiv*, 2024. 5
39. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 8
40. Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE, 2019. 10, 11
41. Rubèn Tito, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2021 competition on document visual question answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16*, pages 635–649. Springer, 2021. 10, 11
42. Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. MLX: Efficient and flexible machine learning on apple silicon, 2023. 12
43. Prince Canuma. MLX-VLM: Mlx-vlm is a package for inference and fine-tuning of vision language models (vlms) on mac using mlx., 2024. 13