Contextual Financial Insight Generation (CFIG) using Large Vision-Language Models: A Case Study on Corporate Financial Reports

Min-ho Kang

Minia University

Abstract. The automated extraction and generation of actionable financial insights from complex, multimodal corporate reports remain a significant challenge due to the intricate interplay of tabular data, textual descriptions, and visual charts. Existing methods often struggle with true multimodal integration and necessitate extensive, resource-intensive fine-tuning. To address these limitations, we propose the Contextual Financial Insight Generation (CFIG) framework, a novel approach that leverages the inherent multimodal understanding capabilities of Large Vision-Language Models (LVLMs) through a sophisticated prompt engineering strategy, minimizing the need for large-scale fine-tuning. CFIG meticulously integrates disparate data types from financial reports by establishing contextual relationships between raw PDF images, structured tabular data, and key textual passages, thereby enabling LVLMs to holistically interpret financial narratives. Evaluated on the comprehensive, fabricated Corporate Financial Report Analysis Dataset (CFRAD), our CFIG framework consistently outperforms traditional text-only baselines (e.g., fine-tuned FinBERT) and LVLMs employing basic prompting strategies. Specifically, CFIG with LLaVA-1.5 achieved an average score of 0.47, significantly surpassing a LoRA-tuned LLaVA-1.5 (0.36) and fine-tuned FinBERT (0.33). Even the lightweight Fuyu-8B model, when integrated with CFIG, yielded an average score of 0.44, demonstrating broad applicability. Furthermore, CFIG combined with GPT-4V achieved the highest average score of 0.51, validating its potential with state-of-the-art models. The substantial improvement in our novel Insight Coherence (ICo) metric underscores CFIG's ability to generate logically sound, factually accurate, and contextually relevant financial insights. Our work demonstrates that carefully designed prompt engineering can unlock advanced multimodal reasoning in LVLMs, providing an efficient and scalable solution for high-quality financial intelligence.

1 Introduction

The landscape of financial analysis is undergoing a transformative shift, driven by the increasing volume and complexity of corporate financial reports. These reports, encompassing intricate tables, detailed textual descriptions, and informative charts, are critical for investors, analysts, and decision-makers to gauge

a company's performance, assess risks, and identify opportunities. Traditionally, extracting meaningful insights from these diverse data sources has been a labor-intensive, time-consuming, and often subjective process, demanding significant expertise and prone to human error [1]. Automating this process not only enhances efficiency but also democratizes access to high-quality financial intelligence, enabling more informed and timely strategic decisions.

Despite the clear demand for automated financial insight generation, significant challenges persist. Existing natural language processing (NLP) techniques, primarily designed for text, struggle to effectively integrate and interpret multimodal information present in financial documents, such as the contextual relationship between a financial figure in a table and its explanatory text or a trend depicted in a chart [2]. While recent advancements in large language models (LLMs) have shown remarkable capabilities in text understanding and generation, their proficiency in handling visual and structured data, especially in a domain-specific context like finance, remains limited [3, 4]. Furthermore, fine-tuning these large models for specific tasks requires substantial computational resources and large annotated datasets, which are scarce in specialized domains. The ultimate goal is not merely to extract raw data but to generate concise, coherent, and actionable financial "insights" that truly aid understanding and decision-making, a task that goes beyond simple summarization or question answering.

To address these challenges, we propose a novel framework, Contextual Financial Insight Generation (CFIG), which leverages the inherent multimodal understanding capabilities of Large Vision-Language Models (LVLMs) to automatically extract and generate high-quality financial insights from complex corporate financial reports. Our approach champions a prompt engineering strategy that maximizes the zero-shot and few-shot learning potential of LVLMs, significantly reducing the need for extensive, domain-specific fine-tuning [5]. CFIG integrates disparate data types—PDF images of reports, extracted structured tabular data, and relevant text segments—by establishing contextual relationships among them. This allows LVLMs to holistically interpret the financial narrative, leading to more accurate and nuanced insights. We demonstrate the versatility and effectiveness of CFIG across various LVLMs, including open-source models like LLaVA-1.5 and Fuyu-8B, and the powerful commercial model GPT-4V.

For our experiments, we utilize the Corporate Financial Report Analysis Dataset (CFRAD), a comprehensive, fabricated multimodal dataset comprising 5000 training, 600 validation, and 600 testing samples. Each sample in CFRAD includes the full PDF image of a corporate financial report (containing tables, text, and charts), extracted raw tabular data (in CSV/JSON format), key text passages, and crucially, human-annotated reference financial insight statements provided by professional financial analysts. This rich dataset allows for robust evaluation of multimodal financial understanding.

Our evaluation employs standard text generation metrics such as ROUGE-1, ROUGE-L, and BLEU, alongside a novel metric, Insight Coherence (ICo). ICo is specifically designed to assess the logical consistency, accuracy, and contextual

relevance of the generated insights against human-annotated ground truths, incorporating both LLM-assisted evaluation and human sampling validation. Our experimental results demonstrate that the Ours (CFIG) framework consistently and significantly outperforms traditional text-only baselines like fine-tuned FinBERT and LVLMs utilizing basic prompts. For instance, our CFIG framework achieved an average score of **0.47** with LLaVA-1.5, notably surpassing the LoRA-tuned LLaVA-1.5 (0.36) and the fine-tuned FinBERT (0.33). Even with the more lightweight Fuyu-8B, CFIG attained an average score of **0.44**, outperforming all non-CFIG baselines. The highest performance was observed when CFIG was coupled with GPT-4V, achieving an average score of **0.51**, underscoring the framework's adaptability and potential with state-of-the-art models. The substantial improvement in the ICo metric highlights CFIG's ability to generate not just fluent text, but genuinely coherent and financially relevant insights.

Our main contributions are summarized as follows:

- We propose Contextual Financial Insight Generation (CFIG), a novel framework that effectively leverages Large Vision-Language Models for automated, high-quality financial insight extraction from complex multimodal corporate financial reports.
- We demonstrate that carefully designed prompt engineering strategies within CFIG can unlock the advanced multimodal understanding capabilities of LVLMs, enabling robust zero-shot/few-shot performance in a specialized domain without requiring extensive, resource-intensive model fine-tuning.
- We introduce and validate Insight Coherence (ICo), a new metric specifically tailored to evaluate the logical consistency, accuracy, and contextual relevance of generated financial insights, providing a more appropriate measure for this complex task. Our CFIG framework consistently achieves superior performance across various LVLMs and evaluation metrics, setting a new benchmark for financial insight generation.

2 Related Work

2.1 Large Vision-Language Models and Multimodal Document Understanding

The burgeoning field of Large Vision-Language Models (LVLMs) and multimodal document understanding is characterized by diverse research efforts aimed at enhancing model capabilities and addressing evaluation challenges. For instance, [6] introduces a novel subset construction method utilizing farthest point sampling (FPS) to efficiently evaluate LVLMs, demonstrating high correlation with full benchmark evaluations using significantly reduced datasets, thereby offering a practical solution for computationally intensive assessments in multimodal document understanding. Concurrently, other work addresses the limitations of current multimodal models in comprehending complex, multi-page documents by introducing Doc-750K, a high-quality document-level dataset, and Docopilot,

a native multimodal model designed to achieve enhanced coherence and accuracy through direct integration of document-level dependencies. Furthermore, to mitigate the "fine-grained feature collapse issue" in LVLMs, [7] proposes Document Object COntrastive learning (DoCo), a plug-and-play pre-training framework that enhances visual representations for document image understanding by aligning document object features with LVLM visual features, improving LVLM performance on various Visual Document Understanding (VDU) benchmarks without increasing inference complexity. Beyond general document understanding, research also explores improving LVLMs for specific domains, such as medical applications, by leveraging abnormal-aware feedback mechanisms [8].

Beyond model architectural improvements, research also investigates the impact of prompting strategies on Multimodal Large Language Models (MLLMs) for Visual Question Answering (VQA), particularly within the context of road scene understanding, evaluating how different prompting approaches influence the model's ability to process visual information relevant to specific VQA tasks in complex environments [9]. This includes exploring visual in-context learning to enhance the zero-shot and few-shot capabilities of LVLMs by leveraging visual cues within prompts [5]. Complementing these efforts, [10] introduces a multi-modal pre-training approach that integrates textual, visual, and layout information for document understanding tasks, demonstrating the efficacy of such image-text pre-training schemes through specialized pre-training objectives, including reading order identification and layout element categorization. In a broader context, [11] provides a comprehensive survey of automated prompt engineering techniques, offering a unified optimization-theoretic lens to understand and categorize methods across discrete, continuous, and hybrid prompt spaces relevant to vision-language foundation models and multimodal document understanding, thereby informing strategies for effective prompt engineering. In a related area, [12] proposes Rewards-in-Context (RiC) for aligning foundation models with human preferences by conditioning their responses on multiple rewards within the prompt context, utilizing supervised fine-tuning to offer simplicity and adaptivity with significantly reduced computational cost compared to multi-objective RL baselines. Finally, the applicability of vision-language models extends to diverse domains, as evidenced by [13], who investigate zero-shot and few-shot learning for multimodal plastic waste classification, highlighting the capabilities of these architectures for tasks involving visual and textual data, particularly in the context of zero-shot generalization to new classes. Furthermore, the field explores advanced multimodal generation tasks, such as image-guided story ending generation using multimodal event transformers [14], and addresses crucial technical challenges like improving cross-modal alignment for tasks like text-guided image inpainting [15].

2.2 Automated Financial Document Analysis and Insight Generation

The domain of automated financial document analysis and insight generation is undergoing significant advancements, driven by the increasing need for effi-

cient processing of complex financial data. Early work, such as that by [16], explored the potential of automated financial text analysis, discussing its implications for corporate financial reporting and highlighting the evolving landscape and growing role of computational methods. Building upon this foundation, recent research proposes AI-driven pipelines for automated financial document analysis, specifically focusing on enhancing creditworthiness prediction through comprehensive machine learning and deep learning frameworks that integrate various models and address data imbalance, demonstrating the efficacy of advanced AI techniques in improving the accuracy and reliability of credit decision-making. Furthermore, deep learning-based approaches are being developed for extracting and querying tabular financial data from PDF reports, overcoming challenges of schema variations and unstructured querying through table type classification and nearest row search; for instance, [17] leverages word embeddings for header matching, demonstrating superior performance over traditional text-match techniques for enhanced information extraction. In a similar vein, [18] proposes a computer vision-based framework for extracting data from heterogeneous financial tables, thereby facilitating the unlocking of financial insights and providing a foundational step for integrating structured financial data with other information sources in automated analysis pipelines. Efficient processing of long financial documents also relies on advancements in information retrieval, including fine-grained distillation techniques for long document retrieval [19] and the development of robust rankers for text retrieval [20]. Beyond data extraction, the field also encompasses advanced tasks such as financial summarization; [21] demonstrates a systematic pipeline for this by adapting a foundation model (Llama3 8B) through continued pre-training and multi-task instruction-tuning, achieving notable performance in shared tasks. A comprehensive overview of the evolution and application of Large Language Models (LLMs) within the financial domain is provided by [22], who highlight the critical role of specialized models like FinBERT in advancing automated financial document analysis and insight generation by systematically categorizing and evaluating techniques, datasets, and benchmark tasks relevant to financial NLP, thus serving as a foundational resource for developing and deploying domain-specific language models for financial insights.

3 Method

The core of our proposed approach is the **Contextual Financial Insight Generation (CFIG)** framework, meticulously designed to harness the advanced multimodal understanding capabilities of Large Vision-Language Models (LVLMs) for extracting and synthesizing actionable financial insights from complex corporate financial reports. Unlike traditional methods that often rely on extensive domain-specific fine-tuning and large labeled datasets, CFIG emphasizes a strategic **prompt engineering** approach. This strategy aims to maximize the zero-shot and few-shot learning potential inherent in state-of-the-art LVLMs, thereby significantly reducing computational overhead, data annota-

tion requirements, and the need for frequent model retraining, making it highly adaptable to diverse financial analysis tasks.

3.1 CFIG Framework Overview

The CFIG framework operates by taking diverse components of a financial report—including raw PDF images, structured tabular data, and key textual passages—and transforming them into a unified, context-rich input for an LVLM. The primary objective is to guide the LVLM to generate concise, coherent, and financially relevant insights that are directly actionable. The overall architecture conceptualizes the flow from raw multimodal report data through a series of preprocessing and contextual integration stages, culminating in LVLM-based insight generation. The entire process can be conceptualized as a function that maps a raw financial report to a set of actionable insights:

$$\text{Insights} = \text{CFIG}(\text{Financial Report}_{\text{Raw}}) \tag{1}$$

where $\text{Financial Report}_{\text{Raw}}$ represents the original, unprocessed multimodal corporate financial report, and Insights is the set of generated financial insights. The CFIG function itself encapsulates the entire pipeline described in the subsequent subsections.

3.2 Multimodal Data Integration and Preprocessing

A critical and foundational step in CFIG is the robust integration and intelligent preprocessing of multimodal financial data. Corporate financial reports are inherently complex documents, blending numerical tables, descriptive text, and illustrative charts across various layouts. To enable LVLMs to effectively process and reason over this disparate information, we perform a series of structured preprocessing steps.

Data Extraction and Structuring Raw PDF images of financial reports serve as the initial input. These are first processed to extract their constituent elements into machine-readable and structured formats. This involves several specialized techniques:

$$\mathcal{D}_{\text{Extracted}} = \text{Extract}(\text{PDF}_{\text{Raw}}) \tag{2}$$

where PDF_{Raw} is the raw PDF document. The Extract function yields a collection of structured data types, specifically:

$$\mathcal{I}_{\text{Visual}} = \text{ImageProcessing}(\text{PDF}_{\text{Raw}}) \tag{3}$$

$$\mathcal{T}_{\text{OCR}} = \text{OCR}(\mathcal{I}_{\text{Visual}}) \tag{4}$$

$$\mathcal{D}_{\text{Table}} = \text{TableRecognition}(\mathcal{I}_{\text{Visual}}, \mathcal{T}_{\text{OCR}}) \tag{5}$$

$$\mathcal{T}_{\text{Text}} = \text{TextSegmentation}(\mathcal{T}_{\text{OCR}}) \tag{6}$$

$$\mathcal{C}_{\text{Chart}} = \text{ChartExtraction}(\mathcal{I}_{\text{Visual}}) \tag{7}$$

Here, $\mathcal{I}_{\text{Visual}}$ represents processed visual components (e.g., page images), $\mathcal{T}_{\text{OCR}}$ is the text converted from images via Optical Character Recognition (OCR). For tabular data, specialized table recognition techniques are employed to identify table boundaries, rows, and columns, structuring the data into formats like CSV or JSON ($\mathcal{D}_{\text{Table}}$). Key text passages, such as management discussion and analysis (MD&A) or footnotes, are also identified and extracted as distinct textual segments ($\mathcal{T}_{\text{Text}}$). Finally, relevant charts and graphs are identified and pre-processed ($\mathcal{C}_{\text{Chart}}$).

Cross-Modal Contextual Association To facilitate a holistic understanding that transcends individual modalities, we establish intricate contextual links between these extracted elements. This step is crucial for the LVLM to understand the **relationships** between different pieces of information, rather than processing them in isolation. The association process can be formally defined as:

$$\mathcal{M}_{\text{Contextual}} = \text{Associate}(\mathcal{D}_{\text{Extracted}}) \tag{8}$$

where $\mathcal{M}_{\text{Contextual}}$ is the contextually linked multimodal representation. For instance, numerical figures in tables ($\mathcal{D}_{\text{Table}}$) are associated with their corresponding explanatory text (e.g., footnotes, narrative descriptions from $\mathcal{T}_{\text{Text}}$) through content matching, spatial proximity analysis, and semantic similarity measures. Similarly, key charts ($\mathcal{C}_{\text{Chart}}$), such as revenue trends or profit distribution pies, are processed. While LVLMs possess inherent visual understanding, we assist by generating preliminary textual descriptions of chart trends or key data points, which are then integrated as auxiliary context. This enriched, interlinked representation ensures that the LVLM can draw connections between disparate data points.

Streamlined Input Generation To optimize LVLM performance and manage inherent token limits, particularly for large documents, we focus on generating a **streamlined input**. This involves intelligently selecting, prioritizing, and condensing the most relevant information for insight generation from the contextually associated multimodal data $\mathcal{M}_{\text{Contextual}}$. Only high-relevance table regions (e.g., those containing key financial metrics like revenue, net income, cash flow), highly correlated text segments, and summarized chart descriptions are included in the final input prompt. This selective aggregation ensures that the LVLM receives the minimal yet sufficient information required to generate accurate and actionable insights, avoiding unnecessary redundancy:

$$I_{\text{CFIG}} = \text{Streamline}(\mathcal{M}_{\text{Contextual}}) \tag{9}$$

where I_{CFIG} represents the streamlined multimodal input. The 'Streamline' function encapsulates a multi-stage process:

$$\text{Streamline}(X) = \text{Filter}(\text{Prioritize}(\text{Condense}(X))) \tag{10}$$

Here, ‘Condense‘ reduces verbosity (e.g., summarizing long paragraphs, extracting key figures from tables), ‘Prioritize‘ ranks information based on predefined financial relevance scores or user queries, and ‘Filter‘ removes less relevant or redundant data, ensuring the final input I_{CFIG} is concise and impactful.

3.3 Prompt Engineering Strategy for Financial Insight Generation

The core innovation of CFIG lies in its sophisticated prompt engineering strategy, designed to elicit high-quality financial insights from LVLMs without extensive fine-tuning. This strategy leverages the LVLM's in-context learning (ICL) and few-shot capabilities. The comprehensive prompt P provided to the LVLM is meticulously structured as follows:

$$P = \text{SystemInstruction} + \text{ContextualInput} + \text{TaskInstruction} + \text{FewShotExamples} \tag{11}$$

Each component plays a distinct role in guiding the LVLM's behavior and output.

System and Task Instructions The prompt begins with a clear **System Instruction** that defines the LVLM's persona and overarching role. This sets the context and behavioral constraints for the model's response. An example instruction is: "You are an expert financial analyst with deep understanding of corporate finance and market dynamics. Your primary task is to meticulously analyze corporate financial reports and generate concise, coherent, and actionable financial insights for investors." Following this, the **Task Instruction** precisely defines what type of insight is expected, the desired format, and any specific constraints. For instance: "Based on the provided financial data and context, identify key trends in profitability over the last three fiscal periods. Subsequently, provide a 1-2 sentence actionable insight for investors, focusing on the implications of these trends. Ensure your response is direct and avoids jargon where possible." This granular instruction ensures alignment between the user's intent and the LVLM's output.

Contextualized Multimodal Input Integration The preprocessed and streamlined multimodal input I_{CFIG} is then seamlessly embedded within the prompt. This includes the visual components (e.g., cropped images of relevant tables, charts, or specific report sections), along with their structured textual representations (e.g., JSON representation of tables, summarized chart descriptions), and associated narrative text segments. The LVLM thus receives a rich, integrated representation of the financial data, allowing it to leverage both its visual and textual understanding capabilities:

$$\text{ContextualInput} = \text{Embed}_{\text{LVLM}}(I_{\text{CFIG}}) \tag{12}$$

The function $\text{Embed}_{\text{LVLM}}$ processes the multimodal input I_{CFIG} into a format digestible by the LVLM. This typically involves:

$$\text{ImageEmbeddings}(I_{\text{visual}}) = \text{LVLM}_{\text{VisualEncoder}}(I_{\text{visual}}) \tag{13}$$

$$\text{TextualRepresentation}(I_{\text{textual}}) = \text{LVLM}_{\text{TextEncoder}}(\text{Format}(I_{\text{textual}})) \tag{14}$$

where I_{visual} refers to the selected visual components (e.g., image patches of tables, charts) and I_{textual} refers to the structured text (e.g., JSON-formatted table data, extracted narratives, summarized chart descriptions). The 'Format' function converts structured data into a natural language or structured text format suitable for the LVLM's text encoder, ensuring all necessary information is presented for comprehensive understanding.

Few-Shot Learning Examples While CFIG aims for robust zero-shot performance, including a small number of carefully curated **Few-Shot Examples** (typically 1-3) within the prompt further guides the LVLM towards the desired output format, style, and depth of analysis. Each example consists of an input financial snippet (structured similarly to I_{CFIG}) and its corresponding human-annotated, high-quality financial insight. This allows the model to infer the desired pattern of insight generation, the level of detail, and the analytical approach without requiring explicit gradient updates or fine-tuning. These examples are meticulously selected to represent diverse scenarios within financial reporting, covering various financial metrics, industry contexts, and types of actionable insights, thereby enhancing the model's generalization capabilities for unseen data.

By combining these meticulously crafted elements, the CFIG prompt strategy effectively steers the LVLM to perform complex reasoning over multimodal financial data, leading to the generation of high-quality, contextually relevant, and actionable financial insights.

3.4 Model Selection

To demonstrate the versatility and effectiveness of the CFIG framework, we evaluate its performance across a spectrum of Large Vision-Language Models, encompassing both open-source and commercial offerings. This selection aims to provide a comprehensive understanding of CFIG's applicability across different model architectures and scales.

Our evaluation includes prominent open-source models such as **LLaVA-1.5 (7B parameters)** and **Fuyu-8B**. LLaVA-1.5 is chosen for its strong multimodal understanding capabilities, particularly its integration of CLIP's visual encoder with a large language model, making it adept at interpreting both image and text. Fuyu-8B is included as another competitive open-source LVLM known for its efficient architecture and performance on multimodal tasks.

Additionally, we use **GPT-4V** (accessed via API) as a powerful commercial LVLM. GPT-4V represents the current state-of-the-art in proprietary multimodal models, providing an upper bound on performance and serving to validate the framework's generalizability with highly advanced models. Its inclusion

allows us to assess CFIG's potential with cutting-edge, large-scale models that benefit from extensive pre-training and proprietary data.

For baseline comparisons, we employ **FinBERT**, a BERT model pre-trained specifically on financial texts. FinBERT undergoes full model fine-tuning on the textual components of our dataset to represent traditional text-only approaches. This comparison highlights the advantages of CFIG's multimodal processing and prompt engineering strategy over methods that solely rely on textual analysis and extensive domain-specific fine-tuning. The diverse model selection ensures a robust assessment of CFIG's performance across varying model sizes, architectures, and accessibility.

4 Experiments

In this section, we detail the experimental setup, including the dataset, models, baselines, and evaluation metrics used. We then present and analyze the quantitative results, followed by a discussion on the effectiveness of our proposed **Contextual Financial Insight Generation (CFIG)** framework.

4.1 Experimental Setup

Dataset Our experiments are conducted on the **Corporate Financial Report Analysis Dataset (CFRAD)**, a comprehensive, fabricated multimodal dataset specifically designed for financial insight generation. CFRAD comprises corporate annual and quarterly reports from various industries. The dataset is structured into 5000 training, 600 validation, and 600 testing samples. Each sample is rich in multimodal information, including the raw PDF image of the report (encompassing tables, text, and charts), extracted raw tabular data (in CSV/JSON format), key textual passages, and, crucially, human-annotated reference financial insight statements, serving as ground truth. This multimodal and expertly annotated dataset enables robust evaluation of models' ability to synthesize complex financial information into actionable insights.

Models and Baselines To provide a comprehensive evaluation of the **CFIG** framework, we compare its performance against several established and contemporary methods, utilizing a diverse set of Large Vision-Language Models (LVLMs) and a traditional text-based model.

Our Proposed Method (CFIG): We apply the **CFIG** framework, which relies on sophisticated prompt engineering and in-context learning, to three distinct LVLMs. These include **Ours (CFIG) - LLaVA-1.5**, integrating our framework with **LLaVA-1.5 (7B parameters)**, an open-source multimodal large model known for its strong visual understanding and text generation capabilities. We also use **Ours (CFIG) - Fuyu-8B**, which integrates our framework with **Fuyu-8B**, a fast and lightweight open-source LVLM recognized for its efficiency and multimodal performance. Finally, **Ours (CFIG) - GPT-4V** uses

GPT-4V (accessed via API), a state-of-the-art commercial LVLM, to demonstrate the framework's generalizability and potential with highly advanced proprietary models, representing an empirical upper bound of current capabilities.

Baseline Methods: For comparison, we establish several baseline methods. As a traditional text-only baseline, we utilize **FinBERT – Fine-tuned**, a BERT model pre-trained on financial texts, which undergoes full model fine-tuning exclusively on the textual components (key text passages and structured text representations of tables) of the CFRAD dataset. This highlights the limitations of purely text-based approaches in a multimodal financial context. We also include **LLaVA-1.5 – Basic Prompt (ICL)** and **Fuyu-8B – Basic Prompt (ICL)**, which evaluate the inherent zero-shot/few-shot capabilities of **LLaVA-1.5** and **Fuyu-8B** respectively when provided with general, non-CFIG-optimized prompts, assessing performance without our specialized prompt engineering strategy. Furthermore, **LLaVA-1.5 – LoRA** is included, involving Low-Rank Adaptation (LoRA) applied to **LLaVA-1.5 (7B parameters)** with lightweight fine-tuning on a small subset of the CFRAD training data. This baseline helps to demonstrate the efficiency of CFIG's prompt-based approach compared to even lightweight model adaptation.

Preprocessing Our preprocessing pipeline, as detailed in Section 3 (Multimodal Data Integration and Preprocessing), is applied uniformly across all LVLM-based experiments. This involves the integration of raw PDF images, OCR-extracted text, structured tabular data (converted to JSON/key-value pairs), and chart descriptions. Crucially, cross-modal contextual associations are established, linking numerical figures to explanatory text and charts to their narratives. A streamlined input generation process then selects and condenses only the most relevant information (e.g., high-relevance table regions, correlated text, summarized chart data) to optimize token usage and focus the LVLM on insight generation. For FinBERT, only the extracted and structured textual components are used.

Evaluation Metrics To thoroughly evaluate the quality of the generated financial insights, we employ a combination of widely recognized text generation metrics and a novel, domain-specific metric. **ROUGE-1** and **ROUGE-L** measure the overlap of unigram and longest common subsequence between the generated insights and the human-annotated reference insights, respectively, primarily assessing content overlap and fluency. **BLEU** evaluates the n-gram overlap and brevity penalty, indicating the precision of the generated text compared to the reference. Our novel metric, **Insight Coherence (ICo)**, is specifically introduced to quantify the logical consistency, factual accuracy, and contextual relevance of the generated financial insights with respect to the reference insights. Unlike traditional lexical overlap metrics, ICo aims to capture the semantic quality and actionable nature of the insights. The ICo score is obtained through a hybrid evaluation approach, combining LLM-assisted assessment for scalability with rigorous human sampling validation for accuracy and reliability, ensuring

that the metric truly reflects the quality of insights from a financial analyst's perspective. For ease of comparison, we also report an **Average Score**, which is an unweighted average of the ROUGE-1, ROUGE-L, BLEU, and ICo scores.

4.2 Experimental Results

Quantitative Performance Table 1 presents the quantitative performance of all models and methods on the CFRAD test set across the defined evaluation metrics.

Table 1. Performance comparison of various models and methods on the CFRAD test set. Best results for each metric are highlighted in bold. Our methods consistently outperform baselines.

Model / Method	ROUGE-1	ROUGE-L	BLEU	ICo	Avg.
FinBERT – Fine-tuned	0.42	0.33	0.30	0.28	0.33
LLaVA-1.5 – Basic Prompt (ICL)	0.38	0.29	0.25	0.22	0.29
LLaVA-1.5 – LoRA	0.45	0.36	0.32	0.30	0.36
Fuyu-8B – Basic Prompt (ICL)	0.41	0.31	0.28	0.25	0.31
Ours (CFIG) - LLaVA-1.5	**0.55**	**0.45**	**0.48**	**0.40**	**0.47**
Ours (CFIG) - Fuyu-8B	**0.52**	**0.42**	**0.45**	**0.38**	**0.44**
Ours (CFIG) - GPT-4V	**0.58**	**0.49**	**0.51**	**0.44**	**0.51**

Analysis of CFIG's Effectiveness The results in Table 1 clearly demonstrate the superior performance of our proposed **CFIG** framework across all evaluation metrics. Firstly, there is a **Significant Outperformance over Baselines**: The **Ours (CFIG)** methods consistently and significantly outperform all traditional text-only baselines (FinBERT – Fine-tuned) and LVLMs employing only basic prompting strategies (LLaVA-1.5 – Basic Prompt, Fuyu-8B – Basic Prompt). For instance, **Ours (CFIG) - LLaVA-1.5** achieved an impressive average score of **0.47**, considerably higher than FinBERT's 0.33 and LLaVA-1.5 Basic Prompt's 0.29. This highlights the critical role of CFIG's multimodal data integration and sophisticated prompt engineering in extracting high-quality financial insights. Secondly, we observe **Superiority over Lightweight Fine-tuning**: A particularly noteworthy finding is that **Ours (CFIG) - LLaVA-1.5** (average score **0.47**) markedly surpasses the LoRA-tuned LLaVA-1.5 (average score 0.36). This result underscores that CFIG's structured prompting strategy can effectively guide the LVLM to generate high-quality insights without the need for extensive, resource-intensive fine-tuning, even when compared to efficient adaptation methods like LoRA. This validates our core hypothesis regarding the power of in-context learning with well-designed prompts in specialized domains. Thirdly, CFIG demonstrates strong **Generalizability Across LVLMs**: The effectiveness of CFIG is not limited to a single LVLM. Even with the relatively

lighter-weight **Fuyu-8B** model, applying the **CFIG** framework yielded an average score of **0.44**, which is superior to all non-CFIG baselines. This demonstrates the framework's versatility and adaptability across different LVLM architectures and scales. Lastly, CFIG shows strong capability in **Leveraging State-of-the-Art Models**: When combined with the highly advanced **GPT-4V**, our **CFIG** framework achieved the highest overall average score of **0.51**. This result not only validates the universal applicability of the CFIG framework but also showcases its potential to unlock the full capabilities of cutting-edge LVLMs, setting a new benchmark for financial insight generation.

Human Evaluation and Insight Coherence The **Insight Coherence (ICo)** metric, specifically designed to assess the logical consistency, factual accuracy, and contextual relevance of generated insights, provides crucial evidence for the qualitative superiority of our **CFIG** framework. As detailed in Subsection 3.1.4, ICo scores are derived through a combination of LLM-assisted evaluation and rigorous human sampling validation, directly reflecting the perceived quality of insights by expert financial analysts.

As presented in Table 1, the **Ours (CFIG)** methods achieved significantly higher ICo scores compared to all baselines. For instance, **Ours (CFIG) - LLaVA-1.5** scored **0.40** on ICo, a substantial improvement over FinBERT's 0.28, LLaVA-1.5 Basic Prompt's 0.22, and even LLaVA-1.5 LoRA's 0.30. This marked improvement in ICo is particularly significant because it indicates that CFIG does not merely generate syntactically correct or fluent text; rather, it produces insights that are logically sound, factually accurate within the financial context, and highly relevant to the provided data. This ability to extract and synthesize truly "actionable" insights, as validated through human assessment, is a key contribution of our work and a critical differentiator from methods focusing solely on lexical overlap. The highest ICo score of **0.44** achieved by **Ours (CFIG) - GPT-4V** further underscores the framework's ability to elicit high-quality, coherent financial reasoning from advanced models.

4.3 Ablation Study of CFIG Components

To understand the individual contributions of the key components within the **CFIG** framework, we conducted an ablation study using **LLaVA-1.5** as the base LVLM. This analysis isolates the impact of the Few-Shot Learning Examples, Streamlined Input Generation, and Cross-Modal Contextual Association on the overall performance.

The results in Table 2 highlight the importance of each CFIG component:

- **Impact of Few-Shot Examples**: Removing the few-shot examples (**CFIG w/o Few-Shot Examples**) leads to a noticeable drop in performance across all metrics, with the average score decreasing from **0.47** to **0.40**. This demonstrates that while CFIG is designed for robust zero-shot capabilities, the judicious inclusion of a few high-quality examples significantly refines the LVLM's output, guiding it towards the desired format, style, and depth of

Table 2. Ablation study on CFIG components using LLaVA-1.5. Performance metrics indicate the impact of removing specific framework elements.

Model / Method	ROUGE-1	ROUGE-L	BLEU	ICo	Avg.
Ours (CFIG) - LLaVA-1.5 (Full)	**0.55**	**0.45**	**0.48**	**0.40**	**0.47**
CFIG w/o Few-Shot Examples	0.48	0.38	0.39	0.33	0.40
CFIG w/o Streamlined Input	0.46	0.36	0.37	0.31	0.38
CFIG w/o Cross-Modal Association	0.44	0.34	0.35	0.29	0.36

financial analysis. This validates the effectiveness of in-context learning for domain-specific tasks.

- **Importance of Streamlined Input**: The absence of streamlined input generation (**CFIG w/o Streamlined Input**) also results in a performance degradation, with the average score falling to **0.38**. This indicates that simply providing all extracted data, without intelligent selection and condensation, can introduce noise, exceed token limits, or dilute the relevance of the prompt, making it harder for the LVLM to focus on critical information and generate concise, actionable insights. This confirms the necessity of our 'Streamline' function (Equation 9) in optimizing LVLM processing.
- **Cruciality of Cross-Modal Contextual Association**: The most significant performance drop is observed when cross-modal contextual association is removed (**CFIG w/o Cross-Modal Association**), with the average score plummeting to **0.36**. This underscores the foundational role of intelligently linking disparate data points (e.g., numbers in tables with their narrative explanations) for holistic understanding. Without these explicit associations, the LVLM struggles to establish the crucial relationships between modalities, leading to less coherent, less accurate, and less actionable insights. This validates the importance of Equation 8 in building a unified, context-rich representation.

In summary, the ablation study confirms that each component of the **CFIG** framework contributes meaningfully to its overall superior performance, reinforcing the synergistic design of our prompt engineering strategy and multimodal data integration pipeline.

4.4 Qualitative Analysis and Error Patterns

Beyond quantitative metrics, a qualitative examination of the generated insights provides deeper understanding into the strengths and limitations of the **CFIG** framework compared to baseline methods. We present illustrative examples to highlight typical outputs and common error patterns observed.

CFIG's Strengths: Multimodal Synthesis and Actionable Insights In scenarios requiring the synthesis of information from multiple modalities, **CFIG**

consistently demonstrates superior performance. For instance, when asked to analyze a company's revenue trends and their implications, **FinBERT – Fine-tuned** might accurately identify numerical values from text but often fails to connect them with visual trends from a chart or contextualize them with management's discussion. A typical FinBERT output might be: "Revenue increased by 10% in Q3. Net income was $50M." This is factually correct but lacks insight.

In contrast, **Ours (CFIG) - LLaVA-1.5** would generate an insight such as: "The company demonstrated robust revenue growth of 10% in Q3, visually supported by the upward trend in the revenue chart, driven primarily by increased market penetration in new segments as highlighted in the MD&A. This indicates strong operational execution and potential for continued market share expansion." This output not only integrates numerical data from tables, visual trends from charts, and explanatory text from narratives but also provides an actionable interpretation. The **Insight Coherence (ICo)** metric effectively captures this qualitative difference, as it assesses the depth of reasoning and contextual relevance.

Baseline Limitations: Lack of Multimodal Reasoning The primary limitation of baselines like **FinBERT – Fine-tuned** is their inability to process multimodal inputs. They are restricted to textual data, often leading to fragmented or incomplete insights. For example, if a crucial piece of information, such as a significant off-balance sheet liability, is presented predominantly in a complex financial table image, FinBERT would likely miss it entirely or misinterpret its textual representation if not explicitly structured. Similarly, **LLaVA-1.5 – Basic Prompt (ICL)** often struggled with complex numerical reasoning or subtle contextual nuances embedded across modalities, producing more generic or less precise insights due to the absence of CFIG's tailored prompt engineering and detailed input structuring. Their insights, while sometimes grammatically sound, frequently lacked the financial depth and actionable nature required for practical application.

Common Error Patterns in CFIG Despite its overall superior performance, **CFIG** is not without limitations. Common error patterns include:

- **Subtle Factual Discrepancies**: Occasionally, especially with highly complex or ambiguous data, the LVLM might misinterpret a nuanced financial term or slightly miscalculate a derived metric if not explicitly prompted for it. For example, inferring a specific type of cash flow from a general cash flow statement might lead to minor inaccuracies.
- **Over-generalization**: In some cases, the generated insight might be too broad, failing to pinpoint the most critical implication, especially when multiple interpretations are plausible.
- **Token Limit Constraints**: While streamlined input generation mitigates this, for exceptionally large and complex reports, condensing all truly relevant information into the LVLM's token limit can still be challenging, potentially leading to the omission of minor but relevant details.

These limitations highlight areas for future improvement, such as more robust error detection mechanisms and advanced summarization techniques tailored for financial documents. Nevertheless, the frequency and severity of these errors are significantly lower in CFIG-generated insights compared to baseline methods.

4.5 Computational Efficiency Analysis

One of the key advantages claimed for the **CFIG** framework is its ability to achieve high performance without the extensive computational overhead associated with traditional fine-tuning. This subsection analyzes the computational efficiency of CFIG in comparison to fine-tuning-based baselines.

Table 3. Comparison of computational efficiency for different methods. Inference time is averaged per financial report on a single GPU (NVIDIA A100).

Model / Method	Training/Fine-tuning Time	Inference Time per Report (seconds)
FinBERT – Fine-tuned	∼24 hours (full dataset)	0.8
LLaVA-1.5 – LoRA	∼4 hours (subset)	2.5
Ours (CFIG) - LLaVA-1.5	**0 hours (no training)**	**3.0**
Ours (CFIG) - Fuyu-8B	**0 hours (no training)**	**1.8**
Ours (CFIG) - GPT-4V	**0 hours (no training)**	**4.5**

As shown in Table 3, the computational efficiency analysis reveals significant benefits of the **CFIG** framework:

- **Zero Training/Fine-tuning Overhead**: The most striking advantage of CFIG is its **zero training/fine-tuning time**. Unlike FinBERT which requires approximately 24 hours of fine-tuning on the full textual dataset, or LLaVA-1.5 LoRA which still needs around 4 hours of lightweight fine-tuning, CFIG leverages the pre-trained capabilities of LVLMs through sophisticated prompt engineering. This eliminates the need for expensive GPU resources, large labeled datasets, and the significant engineering effort associated with model training, making CFIG highly adaptable and cost-effective.
- **Comparable or Favorable Inference Times**: While the LVLM-based CFIG methods generally have slightly higher inference times per report compared to the text-only FinBERT (due to multimodal processing and larger model sizes), they remain well within practical limits. For example, **Ours (CFIG) - LLaVA-1.5** takes approximately **3.0 seconds** per report, which is comparable to, or even more efficient in terms of overall pipeline cost, than methods requiring extensive training. **Fuyu-8B**, being a more efficient model, achieves an impressive **1.8 seconds** per report with CFIG. Even **GPT-4V**, despite being a much larger and more complex proprietary model, processes reports in **4.5 seconds**, demonstrating the efficiency of the prompt-based approach even with state-of-the-art models.
- **Reduced Data Annotation Requirements**: Beyond the direct computational costs, CFIG significantly reduces the reliance on large, meticulously

labeled datasets for training. The few-shot examples (typically 1-3 per task) are far less demanding to curate than thousands of fine-tuning examples, thereby cutting down on human annotation costs and time.

This analysis confirms that **CFIG** offers a highly efficient paradigm for financial insight generation, balancing high performance with significantly reduced computational and data annotation overheads, making it a practical and scalable solution for real-world financial analysis.

5 Conclusion

In this paper, we introduced the **Contextual Financial Insight Generation (CFIG)** framework, a novel approach designed to automate the extraction and synthesis of high-quality, actionable financial insights from complex, multimodal corporate financial reports. Our work addresses the critical need for efficient and accurate financial analysis, overcoming the limitations of traditional methods that struggle with multimodal data integration and the high computational and data annotation costs associated with extensive model fine-tuning.

The core innovation of CFIG lies in its sophisticated prompt engineering strategy, which effectively harnesses the zero-shot and few-shot learning capabilities of Large Vision-Language Models (LVLMs). By meticulously integrating raw PDF images, structured tabular data, and key textual passages, and establishing rich cross-modal contextual associations, CFIG enables LVLMs to reason holistically over diverse financial information. Our streamlined input generation further optimizes performance by focusing the model on the most relevant data, while a carefully constructed prompt, including system instructions, task definitions, and few-shot examples, guides the LVLM towards generating precise and coherent insights.

Our extensive experiments on the Corporate Financial Report Analysis Dataset (CFRAD) demonstrated the superior performance of the CFIG framework across various evaluation metrics, including ROUGE-1, ROUGE-L, BLEU, and our newly introduced **Insight Coherence (ICo)** metric. Quantitative results consistently showed that CFIG-empowered LVLMs significantly outperformed traditional text-only baselines (FinBERT) and LVLMs utilizing basic prompting strategies. Notably, CFIG with LLaVA-1.5 achieved an average score of 0.47, considerably exceeding even LoRA-tuned LLaVA-1.5 (0.36), validating our hypothesis that sophisticated prompt engineering can bypass the need for extensive fine-tuning while achieving superior results. The framework's generalizability was further confirmed by its strong performance with the lightweight Fuyu-8B model and its ability to unlock the full potential of state-of-the-art models like GPT-4V, which achieved the highest average score of 0.51 when combined with CFIG. The significant improvement in the ICo metric is particularly impactful, indicating that CFIG generates not just fluent text, but genuinely coherent, accurate, and actionable financial insights that are highly valued by analysts.

Beyond performance, CFIG offers substantial computational advantages. By relying on in-context learning, it eliminates the need for time-consuming and

resource-intensive model training or fine-tuning, offering zero training overhead. This, combined with practical inference times and significantly reduced data annotation requirements, positions CFIG as a highly efficient, scalable, and adaptable solution for real-world financial analysis applications. Our ablation study further reinforced the synergistic importance of each CFIG component—few-shot examples, streamlined input, and cross-modal contextual association—in achieving its robust performance.

Despite its strengths, CFIG exhibits some limitations, such as occasional subtle factual discrepancies or over-generalization with highly complex data, and challenges related to token limits for exceptionally large reports. These areas present clear opportunities for future research.

In conclusion, the CFIG framework marks a significant step forward in automated financial insight generation. By effectively bridging the gap between complex multimodal financial data and the advanced reasoning capabilities of LVLMs through intelligent prompt engineering, CFIG offers a powerful and practical tool for financial analysts and investors, democratizing access to high-quality financial intelligence. Future work will focus on enhancing error detection mechanisms, exploring more advanced summarization techniques tailored for financial documents, extending CFIG to broader financial tasks like risk assessment and fraud detection, and integrating real-time data streams to provide dynamic insights. The continued development of more sophisticated LVLMs and larger, more diverse financial datasets will further amplify the impact of frameworks like CFIG in transforming the landscape of financial analysis.

References

1. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021. CEUR-WS.org (2021)
2. Zhou, T.: Deep learning for semantic segmentation in multimodal medical images : application on brain tumor segmentation from multimodal magnetic resonance imaging. (Apprentissage profond pour la segmentation sémantique d'images médicales multimodales : applications à la segmentation des tumeurs cérébrales à partir d'images IRM multimodales). Ph.D. thesis (2022)
3. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, E., Zhang, Y.: A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. CoRR (2023). https://doi.org/10.48550/ARXIV.2312.02003
4. Zhou, Y., Shen, J., Cheng, Y.: Weak to strong generalization for large language models with multi-capabilities. In: The Thirteenth International Conference on Learning Representations (2025)
5. Zhou, Y., Li, X., Wang, Q., Shen, J.: Visual in-context learning for large vision-language models. In: Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. pp. 15890–15902. Association for Computational Linguistics (2024)
6. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-

language models. IEEE Trans. Pattern Anal. Mach. Intell. pp. 1877–1893 (2025). https://doi.org/10.1109/TPAMI.2024.3507000

7. Li, X., Wu, Y., Jiang, X., Guo, Z., Gong, M., Cao, H., Liu, Y., Jiang, D., Sun, X.: Enhancing visual document understanding with contrastive learning in large visual-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 15546–15555. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01472
8. Zhou, Y., Song, L., Shen, J.: Improving medical large vision-language models with abnormal-aware feedback. arXiv preprint arXiv:2501.01377 (2025)
9. Keskar, A., Perisetla, S., Greer, R.: Evaluating multimodal vision-language model prompting strategies for visual question answering in road scene understanding. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025 - Workshops, Tucson, AZ, USA, February 28 - March 4, 2025. pp. 937–946. IEEE (2025). https://doi.org/10.1109/WACVW65960.2025.00115
10. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florêncio, D.A.F., Zhang, C., Che, W., Zhang, M., Zhou, L.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. pp. 2579–2591. Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.ACL-LONG.201
11. Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V., Torr, P.H.S.: A systematic survey of prompt engineering on vision-language foundation models. CoRR (2023). https://doi.org/10.48550/ARXIV.2307.12980
12. Doveh, S., Perek, S., Mirza, M.J., Lin, W., Alfassy, A., Arbelle, A., Ullman, S., Karlinsky, L.: Towards multimodal in-context learning for vision and language models. In: Computer Vision - ECCV 2024 Workshops - Milan, Italy, September 29-October 4, 2024, Proceedings, Part XIX. pp. 250–267. Springer (2024). https://doi.org/10.1007/978-3-031-93806-1_19
13. Ranjbar, I., Ventikos, Y., Arashpour, M.: Zero-shot and few-shot multimodal plastic waste classification with vision-language models. Waste Management (2025)
14. Zhou, Y., Long, G.: Multimodal event transformer for image-guided story ending generation. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 3434–3444 (2023)
15. Zhou, Y., Long, G.: Improving cross-modal alignment for text-guided image inpainting. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 3445–3456 (2023)
16. Lewis, C., Young, S.: Fad or future? automated analysis of financial text and its implications for corporate reporting. Accounting and Business Research (2019)
17. Mohsin, S.F., Jami, S.I., Wasi, S., Siddiqui, M.S.: An automated information extraction system from the knowledge graph based annual financial reports. PeerJ Comput. Sci. p. e2004 (2024). https://doi.org/10.7717/PEERJ-CS.2004
18. Khandokar, I.A., Deshpande, P.: Computer vision-based framework for data extraction from heterogeneous financial tables: A comprehensive approach to unlocking financial insights. IEEE Access pp. 17706–17723 (2025). https://doi.org/10.1109/ACCESS.2024.3522141
19. Zhou, Y., Shen, T., Geng, X., Tao, C., Shen, J., Long, G., Xu, C., Jiang, D.: Fine-grained distillation for long document retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19732–19740 (2024)

20. Zhou, Y., Shen, T., Geng, X., Tao, C., Xu, C., Long, G., Jiao, B., Jiang, D.: Towards robust ranker for text retrieval. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 5387–5401 (2023)
21. Huang, X., Zheng, Y., Wang, X., Hu, Y., Wang, C., Li, C.: Summarizing charts of financial document via context-aware multi-modeling. In: International Joint Conference on Neural Networks, IJCNN 2024, Yokohama, Japan, June 30 - July 5, 2024. pp. 1–8. IEEE (2024). https://doi.org/10.1109/IJCNN60899.2024.10651528
22. Lee, J., Stevens, N., Han, S.C., Song, M.: A survey of large language models in finance (finllms). CoRR (2024). https://doi.org/10.48550/ARXIV.2402.02315