ViG-LLM: Enhancing Visual Grounding Capabilities in Closed-Box LLMs for Document Information Extraction without OCR Dependencies

Sudhanshu Bhoi
Amazon, India (sudhbee@amazon.com)

amazon | science

Motivation & Introduction

The Problem:

- Large Language Models (LLMs) excel at document processing but lack **visual grounding** (the ability to locate where information comes from).
- Business-critical tasks (finance, legal, medical) require verified data sources.
- Current solutions rely on external OCR (high cost/latency) or model fine-tuning (impossible for closed-box/proprietary models).

The Solution:

- **ViG-LLM:** A novel framework enabling closed-box LLMs (e.g., Claude, Nova) to generate bounding box coordinates without external OCR.
- **Key Advantage:** Achieves high reasoning capabilities with added explainability and reliability.

Methodology

Framework Architecture:

- **Visual Layout Deconstruction:** Uses a U-Net model to segment document images into blocks and create a grid structure.
- **Multi-Agent System:** Decomposes localization into sequential tasks:
  - *Horizontal Viewport Identification Agent (HVIA): Determines the row span of the target information.*
  - *Vertical Viewport Identification Agent (VVIA): Determines the column span within the identified row.*
  - *Visual Grounding Verification Agent (VGVA): Acts as a "Critic" to validate presence within the viewport.*
- **Human-In-The-Loop (HITL):** Incorporates feedback for bounding box corrections and prompt alignment to handle ambiguities.

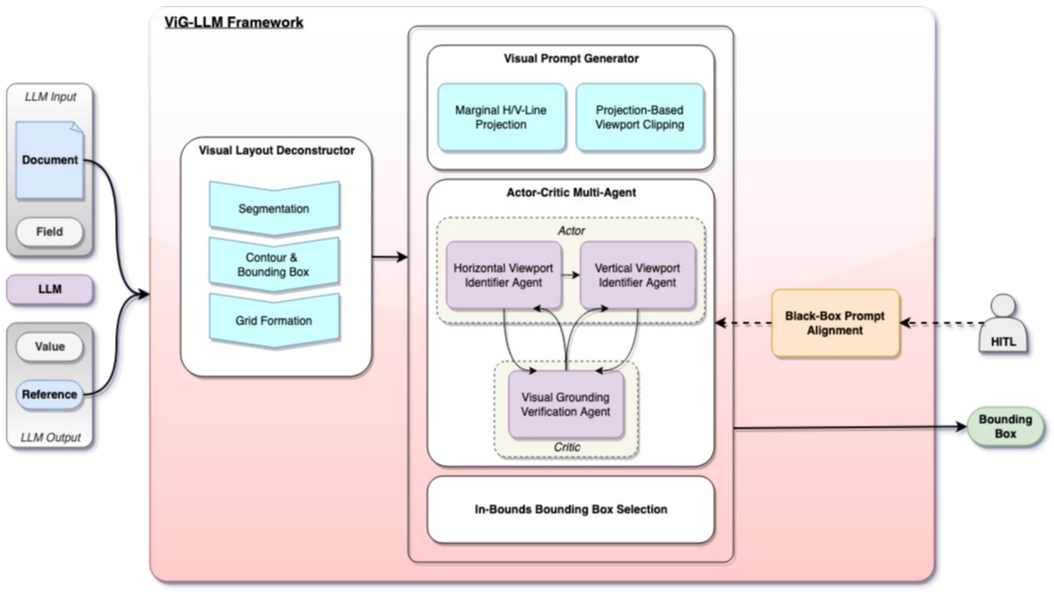

Figure 1: Architecture overview of the ViG-LLM framework. The system comprises three main components: (1) Visual Layout Deconstruction employing U-Net segmentation for grid formation, (2) Multi-Agent LLM system performing viewport identi-fication through visual prompting, and (3) Human-in-the-Loop learning for accuracy refinement. Arrows indicate the flow of information through the system, with dotted lines representing optional feedback paths for continuous improvement.

Figure 2: Sequential stages of viewport identification in the ViG-LLM framework. The progression demonstrates the system's visual grounding process: (1) input document image, (2) segmentation mask generated by the trained U-Net model, (3) horizontal line overlay derived from grid coordinates, (4) horizontal viewport selection with subsequent vertical line overlay in the clipped region, and (5) final bounding box localization

Experimental Results

- **Datasets:** Evaluated on FATURA (10,000 synthetic invoices) and CORD (11,000 receipts).
- **Accuracy:**
  - *Achieved perfect accuracy over spatial-reasoning tuned MLLMs (like Amazon Nova Pro).*
  - *Matches performance of enterprise OCR solutions (e.g., AWS Textract).*
- **Consistency:**
  - *Demonstrated superior template-specific consistency compared to standalone LLMs.*
- **Operational Efficiency:**
  - *Cost: ~$4.16 per 1,000 pages (using Amazon Nova Pro).*
  - *Latency: 3-5 seconds per document (comparable to cloud OCR).*

Approach	Method	Accuracy @ IoU Threshold					
		FATURA			CORD		
		0.25	0.5	0.75	0.25	0.5	0.75
OCR	Amazon Textract (AnalyzeExpense)	**1**	**1**	0.27	**0.96**	**0.96**	**0.96**
	LayoutLMv3 (DocVQA)	0.56	0	0	0.06	0	0
MLLM	Claude Sonnet 4	**0.96**	0.51	0.33	0.31	0.08	0.05
	Claude Opus 4.1	**0.97**	0.7	0.47	0.38	0.14	0.11
	Nova Pro v1	0.8	0.1	0.07	0.27	0.06	0.06
	Nova Premium v1	0.78	0.63	0.09	0.09	0	0
	Qwen2.5-VL-72B-Instruct	**1**	0.63	0.11	0.31	0	0
ViG-LLM (Ours)	Claude Sonnet 4 + ViG-LLM	**0.94**	**0.94**	**0.94**	**0.96**	**0.92**	**0.92**
	Claude Opus 4.1 + ViG-LLM	**1**	**1**	**1**	**0.9**	**0.9**	**0.9**
	Nova Pro v1 + ViG-LLM	**0.91**	**0.91**	**0.91**	0.72	0.67	0.67
	Nova Premier v1 + ViG-LLM	**1**	**1**	**1**	0.88	0.88	0.88

Table 1: Performance Comparison of Document Processing Methods: ViG-LLM Framework demonstrates superior consistency across IoU thresholds, achieves over 90% accuracy for FATURA and CORD datasets with Claude models and significantly outperforms standalone LLM across all architectures.

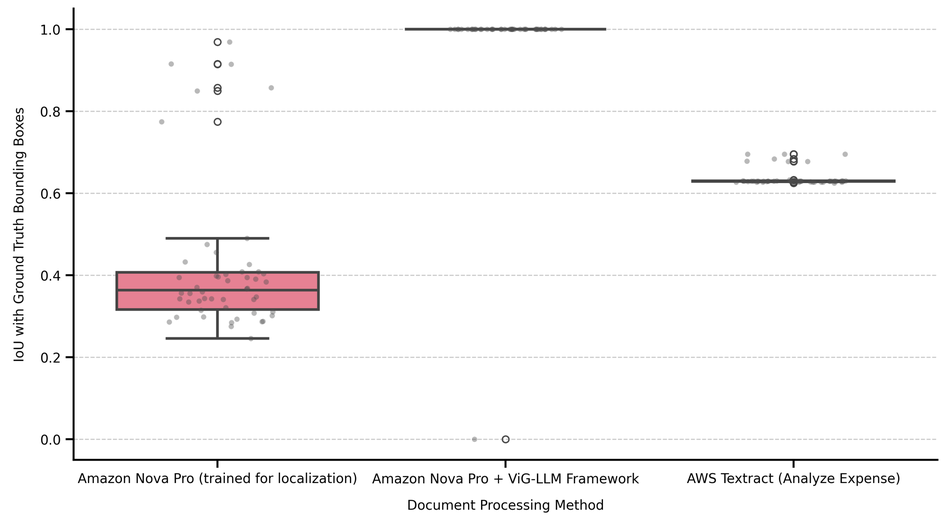

Figure 3: Comparative analysis of template-specific consistency across document processing methods. Performance evaluation depicts Intersection over Union (IoU) scores for (a) standalone Amazon Nova Pro trained for localization, (b) ViG-LLM framework integrated with Amazon Nova Pro and (c) AWS Textract Analyze Expense API. Results demonstrate consistency across multiple instances of identical templates, with box plots indicating median performance, quartile distribution, and outliers for each method. Higher IoU scores and smaller variance indicate superior consistency in visual grounding capabilities as showcased by the ViG-LLM framework.

Conclusion & Future Work

- **Summary:** ViG-LLM successfully decomposes complex localization into manageable viewport tasks using multi-agent collaboration.
- **Impact:** Enables "closed-box" proprietary models to perform visual grounding without expensive OCR dependencies.
- **Future Directions:** Expanding to cross-lingual visual grounding and handling more complex document structures.