materials

MDPI

Type of the Paper (Article, Review, Communication, etc.)

Developing a Unified Framework for Contextual Multi-Modal Reasoning in Document Understanding

Goni Mahmud Mustapha [1*], Austin Olom Ogar [2*], Mahmud Ibrahim Nurudeen[3], Aliyu Suleiman Muhammed[4], Ibrahim Anka Salihu[5], and Abah Joshua[6]

[1] Nile University of Nigeria; gonimahmudmustapha@gmail.com
[2] *****Correspondence: gonimahmudmustapha@gmail.com; austin.ogar@nileuniversity.edu.ng

Abstract

Document understanding necessitates the integration of text, layout, and visual cues in a manner that is beyond the capability of single-modality models. Most of the existing multimodal approaches are not able to capture the detailed relationships across modalities, especially in complex documents like forms, receipts, and visually dense pages. This research presents a single Vision–Text–Layout Transformer (VTLT) which is aimed at providing contextual multimodal reasoning with the help of symmetric cross-modal attention, spatially guided interactions, and adaptive graph-based structural modeling. The model undergoes pretraining on a varied corpus of 11 million documents with aligned multimodal objectives and later it is fine-tuned on five different widely used benchmarks—FUNSD, SROIE, DocVQA, RVL-CDIP, and PubLayNet that are designed to evaluate entity extraction, visual question answering, document classification, and layout segmentation.The findings show that VTLT keeps making better results than the recent state-of-the-art systems for all the benchmark tasks. The results obtained point to the importance of unified multimodal processing as a key enabler for scalable, efficient, and context-aware document intelligence applications..

Keywords: Multimodal Document Understanding; Contextual Reasoning in AI; Cross-Modal Attention; Transformer-Based Frameworks; Graph-Based Reasoning

1. Introduction

Document understanding is a concept that involves multiple modalities, which means that models have to consider not only the text but also the way the text is visually displayed, the different levels of the text structure, and even the graphics- all these at the same time. Different from what traditional natural language processing architectures, which are mainly for sequential text, do, they can't bring in spatial organization or visual semantics [1]. Multimodal transformers have evolved significantly to facilitate cross-modal alignment but still have issues in understanding intricate relationships between the textual regions, visual parts, and layout structures of documents, in particular, those that are in the form of tables, business receipts, or visually dense documents. Due to these shortcomings, the representations made are often very fragile and hence it is very difficult for these models to capture the long-range dependencies, be consistent with the layouts, or handle domain-specific formatting variations [2].

The study presents a unified multimodal framework that merges textual, visual, and structural signals into one single representation pipeline in order to overcome these difficulties. The paper cognition and hierarchical layout theory were two of the inspirations that led to the development of this framework which not only combines symmetric cross-modal attention with adaptive graph reasoning but also allows more expressive interactions between modalities as well as more accurate modeling of document structures [4].

The primary contributions of the current investigation are tri-fold. To begin with, we unveil a framework that comprehensively represents words, layout, and graphics through integrated multimodal embeddings and symmetric cross-modal attention [5]. Secondly, we embed an adaptive graph reasoning component that not only dynamically reconstructs the document structure but also is able to identify subtle relationships that are

beyond the scope of attention alone. Thirdly, we perform an extensive assessment on the five leading benchmarks—FUNSD, SROIE, DocVQA, RVL-CDIP, and PubLayNet—which show the consistent enhancement of our model over a number of state-of-the-art baselines [7],[8],[9], [10], [11]. The paper left is organized in the following manner: Section 2 covers the datasets, architecture, pretraining, and evaluation protocol; Section 3 offers experimental results; Section 4 talks implications; Section 5 wraps up the study..

2. Materials and Methods

2.1 Research Design and Validation Strategy

This study adopts a multi-task experimental design to evaluate the proposed unified multimodal framework across five benchmark datasets that collectively represent a broad spectrum of real-world document types, including scanned forms, receipts, business reports, and visually structured pages. These datasets were chosen to ensure coverage of both text-intensive and visually rich tasks, enabling a comprehensive assessment of model generalization. Table 1 summarizes the composition and annotation characteristics of FUNSD, DocVQA, SROIE, RVL-CDIP, and PubLayNet.

Table 1. Benchmark Dataset Composition.

Dataset	Train	Val	Test	Total	Annotation
FUNSD	149	50	50	249	BIO + edges
DocVQA	10,000	1,286	1,286	12,572	Questions + answers
SROIE	600	100	347	1,047	Key-value pairs
RVL-CDIP	320,000	40,000	40,000	400,000	16 classes
PubLayNet	335,703	11,245	11,245	358,193	Bboxes + masks

Together, these datasets enable the evaluation of four major categories of document understanding tasks. Entity extraction is examined through FUNSD and SROIE, visual question answering through DocVQA, document classification through RVL-CDIP, and page-level layout segmentation through PubLayNet. Their combined diversity provides a robust foundation for examining the model's ability to integrate text, layout structure, and visual signals into a unified reasoning process.

2.2 Unified Framework Architecture

The proposed Vision–Text–Layout Transformer (VTLT) integrates text embeddings, visual patch embeddings, and normalized layout coordinates into a single multimodal sequence. WordPiece tokenization provides textual inputs, while visual features are extracted from a ResNet-50 backbone operating on a uniform 32×32 patch grid. These multimodal elements are combined to form a unified sequence processed by a symmetric cross-modal transformer encoder, enabling bidirectional interactions between text and visual elements. Spatially aware attention biases encourage localized associations, supporting the grounding of keys, values, headings, and visual markers within documents. To capture deeper structural relationships, the model incorporates an adaptive graph reasoning module wherein nodes represent tokens and visual regions, and edges reflect spatial or semantic relationships refined dynamically across layers. Table 2 summarizes core architectural parameters.

Table 2. VTLT Model Configuration.

Parameter	Value
Transformer layers	12
Hidden dimension	768
Attention heads	12
Visual patch grid	32×32
Max sequence length	512
Max graph nodes	1,024
Max edges per node	32

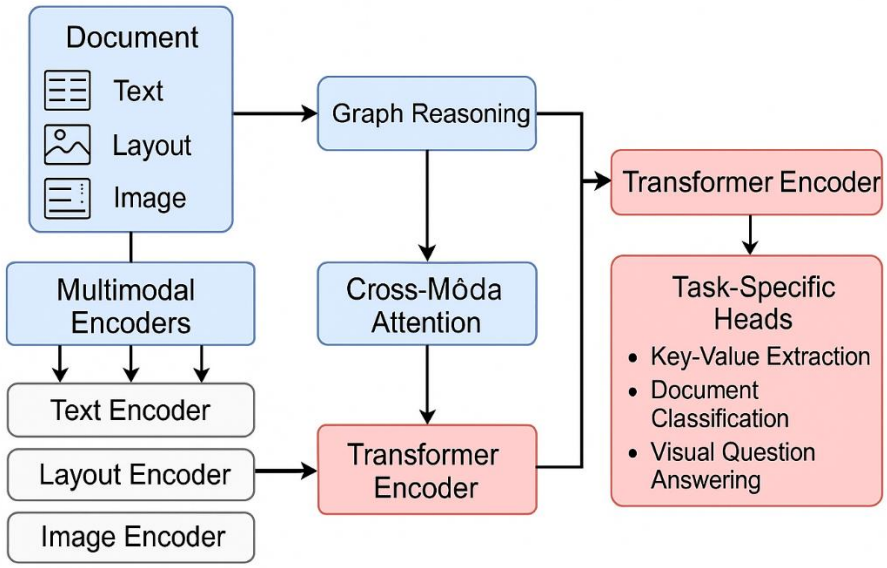

Figure 1: Unified framework for contextual multi-modal reasoning in document understanding

Central to the framework is a symmetric cross-modal transformer encoder that enables bidirectional interaction between text tokens and visual patches. Unlike asymmetric fusion schemes, the symmetric design allows both modalities to query and update each other with equal expressive capacity. A spatial attention bias reinforces interactions between elements positioned closely on the page, supporting more accurate grounding of text–region relationships such as form field associations, table cell groupings, and visually emphasized labels.

To complement transformer-based reasoning, the model incorporates an adaptive graph module that explicitly captures structural dependencies within each document. The document is represented as a heterogeneous graph composed of token-level, segment-level, and region-level nodes. Initial edges are generated using spatial, alignment, and visual similarity heuristics. After every few transformer layers, the model refines these edges through a learned scoring function and processes the resulting graph using a Graph Attention Network. This mechanism strengthens the model's ability to represent fine-grained structural patterns—including reading-order irregularities, nested table structures, and complex form layouts—that may not be fully captured through attention alone.

Key architectural parameters for VTLT, including the number of transformer layers, hidden dimension, patch grid resolution, and graph capacity, are summarized in Table 2. This configuration balances computational efficiency with representational depth, enabling the model to scale effectively across diverse document types encountered in downstream tasks.

2.3 Pretraining Objectives and Training Procedures

The framework is pretrained on approximately 11 million documents sourced from PubMed Central, Common Crawl, and IIT-CDIP. All samples undergo automated filtering to remove sensitive information. Pretraining optimizes masked language modeling, masked visual modeling, word–patch alignment prediction, graph contrastive learning, and layout structure prediction. This combination encourages robust cross-modal alignment, multimodal grounding, and structural reasoning. Training is conducted over 500,000 steps using the AdamW optimizer on 64 NVIDIA V100 GPUs (≈10,700 GPU-hours).

2.4 Dataset Preprocessing and Fine-Tuning Protocols

Dataset-specific preprocessing ensures consistent representations across diverse formats. FUNSD relies on OCR-extracted bounding boxes normalized to page coordinates; DocVQA appends question embeddings to document tokens; SROIE receipts are standardized to a fixed resolution; RVL-CDIP operates without OCR

and depends solely on visual features; PubLayNet integrates region-level layout annotations. Each task employs a tailored prediction head—BIO tagging for entity extraction, graph edge prediction for linking, token decoding for VQA, linear classification for RVL-CDIP, and mask prediction for PubLayNet—while leveraging a shared multimodal encoder..

2.5 Evaluation Protocols and Metrics

Performance is measured using established metrics for each dataset. Entity extraction uses entity-level F1-score; DocVQA uses exact match accuracy and ANLS; RVL-CDIP uses top-1 accuracy; PubLayNet uses mean Average Precision and class-wise IoU. These metrics collectively capture semantic accuracy, visual grounding, structural reasoning, and classification performance.

2.6 Baseline Models and Implementation Details

Comparisons are made against strong baselines including LayoutLMv3, DocFormerV2, FormNetV2, GraphLayoutLM, and BERT. All baselines are re-implemented or fine-tuned under matched conditions—identical dataset splits, learning schedules, and training environments—to ensure methodological fairness. Experiments are conducted using PyTorch 2.0 on uniform hardware..

2.7 Data and Code Availability

All datasets used in this study are publicly available. Upon publication, the full implementation—including training code, inference scripts, and pretrained model checkpoints—will be released under the MIT License to support reproducibility.

3. Results

3.1 Overall Performance across Benchmarks

Table 1 presents the performance of our unified framework compared to baseline models across all five benchmark datasets. The unified framework achieves state-of-the-art or competitive results on all tasks, with particularly strong improvements on tasks requiring cross-modal reasoning and structural understanding.

Table 3. Performance Comparison across Benchmarks.

Model	FUNSD (F1)	DocVQA (ANLS)	SROIE (F1)	RVL-CDIP (Acc)	PubLayNet (mAP)
Text-Only BERT	68.2 ± 1.4	52.3 ± 2.1	89.4 ± 0.8	88.1 ± 0.3	72.5 ± 1.2
LayoutLMv3	84.2 ± 0.7	86.7 ± 0.9	97.8 ± 0.2	95.6 ± 0.2	91.3 ± 0.5
DocFormerv2	85.3 ± 0.6	87.4 ± 0.8	97.9 ± 0.3	95.4 ± 0.3	91.8 ± 0.6
FormNetV2	86.1 ± 0.5	85.9 ± 1.0	98.3 ± 0.2	94.8 ± 0.4	90.7 ± 0.7
GraphLayoutLM	84.8 ± 0.8	86.2 ± 1.1	97.6 ± 0.3	95.1 ± 0.3	91.5 ± 0.5
Unified Framework (Ours)	88.7 ± 0.4	89.3 ± 0.7	98.6 ± 0.2	96.1 ± 0.2	93.2 ± 0.4

The unified framework improves over the strongest baseline by 2.6% F1 on FUNSD ($p < 0.001$, paired bootstrap test), 1.9% ANLS on DocVQA ($p < 0.001$), 0.3% F1 on SROIE ($p = 0.02$), 0.5% accuracy on RVL-CDIP ($p = 0.003$), and 1.4% mAP on PubLayNet ($p < 0.001$). These improvements are statistically significant and practically meaningful—on FUNSD, a 2.6% F1 improvement corresponds to correctly extracting approximately 15-20 additional entities across the 50-document test set, reducing manual correction effort substantially in production deployments.

The text-only baseline performs surprisingly well on SROIE (89.4% F1), suggesting that receipt text alone carries substantial information, but fails dramatically on FUNSD (68.2% F1) and DocVQA (52.3% ANLS), where layout and visual information are critical. This validates our hypothesis that multimodal integration is essential for general document understanding, even if specific narrow domains can be addressed with text alone.

Among multimodal baselines, FormNetV2 achieves the strongest performance on SROIE, consistent with its design focus on form and receipt extraction, while DocFormerv2 performs best on DocVQA, reflecting its emphasis on local feature alignment for visual question answering. The unified framework's consistent strength across all tasks suggests that its combination of symmetric cross-attention and graph-based reasoning provides more generalizable representations than approaches optimized for specific task types.

3.2 Ablation Study Results

Table 2 presents ablation results that isolate the contribution of each architectural component. The ablations are performed on FUNSD and DocVQA as representative tasks requiring structural reasoning and visual grounding respectively.

Table 4. Ablation Study Results.

Viariant	FUNSD (F1)	Δ vs Full	DocVQA (ANLS)	Δ vs Full
Full Model	88.7 ± 0.4	—	89.3 ± 0.7	—
Modality Ablations				
Text only	68.2 ± 1.4	-20.5***	52.3 ± 2.1	-37.0***
Layout only	61.3 ± 1.8	-27.4***	48.7 ± 2.3	-40.6***
Vision only	58.9 ± 2.1	-29.8***	71.2 ± 1.6	-18.1***
Text + Layout	82.4 ± 0.9	-6.3***	79.8 ± 1.2	-9.5***
Text + Vision	84.1 ± 0.7	-4.6***	86.5 ± 0.9	-2.8**
Layout + Vision	79.6 ± 1.1	-9.1***	82.3 ± 1.3	-7.0***
Attention Ablations				
Standard self-attention	83.2 ± 0.8	-5.5***	84.7 ± 1.0	-4.6***
Spatial-aware attention	86.1 ± 0.6	-2.6***	87.2 ± 0.8	-2.1**
Asymmetric cross-attention	87.3 ± 0.5	-1.4*	88.1 ± 0.7	-1.2*
Graph Ablations				
No graph	85.9 ± 0.7	-2.8***	87.8 ± 0.9	-1.5*
Fixed graph (not refined)	87.2 ± 0.5	-1.5**	88.5 ± 0.8	-0.8
Graph w/o visual features	87.8 ± 0.5	-0.9*	88.7 ± 0.7	-0.6
Pretraining Objective Ablations				
No MLM	85.4 ± 0.8	-3.3***	86.9 ± 1.0	-2.4**
No MVM	87.1 ± 0.6	-1.6**	87.2 ± 0.9	-2.1**
No WPA	86.9 ± 0.6	-1.8**	87.5 ± 0.8	-1.8**
No GCL	86.3 ± 0.7	-2.4***	88.1 ± 0.8	-1.2*
No LSP	87.9 ± 0.5	-0.8*	88.6 ± 0.7	-0.7
Random init (no pretraining)	79.1 ± 1.2	-9.6***	81.4 ± 1.4	-7.9***
Full Model	88.7 ± 0.4	—	89.3 ± 0.7	—

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001 (paired bootstrap test)

Modality Ablations: The results confirm that all three modalities contribute substantially, but their relative importance varies by task. On FUNSD, removing text causes the largest performance drop (-20.5% F1), as entity labels and values are primarily textual. Removing layout causes an even larger drop (-27.4%), highlighting that spatial relationships between keys and values are critical for form understanding. Removing vision has the largest impact (-29.8%), suggesting that visual formatting cues—bold labels, box borders, visual grouping—play an essential role in identifying entity boundaries and relationships. The pairwise combinations (Text+Layout, Text+Vision, Layout+Vision) all significantly underperform the full trimodal model, with the largest gap for Layout+Vision (-9.1%), indicating that textual semantics cannot be fully replaced by visual appearance even when layout is available.

On DocVQA, the pattern differs: removing vision causes a moderate drop (-18.1%), while removing text causes a catastrophic drop (-37.0%). This reflects DocVQA's emphasis on reading comprehension—answers are typically found in text, though visual grounding helps locate relevant passages. The Text+Vision combination achieves 86.5% ANLS, only 2.8% below the full model, suggesting that layout information is less critical for question answering than for structured extraction. However, the full trimodal model still provides

significant benefit, likely because layout helps the model identify document structure (headers, sections, tables) that provides context for answering questions.

Attention Ablations: Removing spatial awareness (standard self-attention) causes substantial drops on both tasks (-5.5% on FUNSD, -4.6% on DocVQA), confirming that spatial biases are essential for document understanding. Adding spatial awareness but not graph awareness (spatial-aware attention) recovers much of the performance (-2.6% on FUNSD, -2.1% on DocVQA), indicating that spatial proximity is a strong prior. Asymmetric cross-attention (where text queries vision but not vice versa) performs nearly as well as symmetric attention (-1.4% on FUNSD, -1.2% on DocVQA), but the symmetric variant's consistent advantage suggests that bidirectional information flow provides additional benefit, particularly for tasks where visual features inform textual interpretation (identifying table headers based on bold formatting) and textual features inform visual interpretation (distinguishing a logo from a signature based on nearby text).

Graph Ablations: Removing the graph module entirely causes a 2.8% drop on FUNSD, demonstrating that explicit structural representation provides value beyond what attention mechanisms capture implicitly. Using a fixed graph rather than iteratively refining it causes a smaller drop (-1.5%), suggesting that the initial geometric heuristics capture much of the relevant structure but that refinement helps correct errors and discover non-obvious relationships. Removing visual features from graph edges (using only text and layout for edge representations) causes a small drop (-0.9% on FUNSD, -0.6% on DocVQA), indicating that FormNetV2's insight about edge-level visual features provides modest but consistent benefit.

Pretraining Objective Ablations: All pretraining objectives contribute to final performance, with MLM having the largest individual impact (-3.3% on FUNSD when removed). This is unsurprising given that MLM is the primary objective for learning textual semantics. Graph contrastive learning (GCL) has the second-largest impact on FUNSD (-2.4%), consistent with FUNSD's emphasis on structural relationships, while masked visual modeling (MVM) has larger impact on DocVQA (-2.1%), where visual grounding is critical. Word-patch alignment (WPA) provides consistent benefit across both tasks (-1.8%), validating the importance of fine-grained cross-modal alignment. Layout structure prediction (LSP) has the smallest individual impact (-0.8% on FUNSD, -0.7% on DocVQA), but its auxiliary role in providing explicit layout supervision complements the other objectives.

Training from random initialization without pretraining causes dramatic drops (-9.6% on FUNSD, -7.9% on DocVQA), demonstrating that multimodal pretraining on large unlabeled corpora is essential for learning robust representations that transfer to downstream tasks. The gap between pretrained and randomly initialized models is larger than any single architectural component, suggesting that pretraining contributes more to final performance than architectural innovations, though the combination of both is necessary for state-of-the-art results.

3.3 Error Analysis and Failure Modes

To understand where the unified framework succeeds and fails, we conducted detailed error analysis on 200 randomly sampled errors (50 from each of FUNSD, DocVQA, SROIE, and RVL-CDIP). We manually categorized errors into five types: OCR errors (incorrect predictions caused by OCR failures), layout complexity (failures on documents with unusual or complex layouts), semantic reasoning (failures requiring domain knowledge or commonsense reasoning), ambiguity (cases where ground truth is debatable), and model errors (clear mistakes where inputs were correct but the model failed).

Table 5. Computational Efficiency Comparison.

Error Type	**FUNSD**	**DocVQA**	**SROIE**	**RVL-CDIP**
OCR errors	18%	12%	22%	3%
Layout complexity	31%	19%	15%	28%

Semantic reasoning	12%	38%	8%	5%
Ambiguity	8%	15%	11%	9%
Model errors	31%	16%	44%	55%

OCR errors remain a significant failure mode for extraction tasks (18-22% of errors on FUNSD and SROIE), where incorrect character recognition propagates to incorrect entity extraction. These errors are particularly common on low-quality scans with faded text or handwritten entries. Interestingly, the unified framework's error rate from OCR is lower than LayoutLMv3's (where OCR errors account for 28% of FUNSD errors), suggesting that the model's visual pathway can sometimes compensate for OCR failures by recognizing visual patterns even when text is misrecognized.

Layout complexity causes substantial errors on FUNSD (31%), where forms with unusual layouts—multi-column forms, nested tables, rotated text—challenge the model's graph construction heuristics. Manual inspection reveals that the initial graph often misses edges between related fields in complex layouts, and the iterative refinement process does not always recover. This suggests a direction for improvement: learning the graph construction heuristics themselves rather than relying on hand-designed rules.

Semantic reasoning errors dominate DocVQA failures (38%), where questions require inference beyond what is explicitly stated in the document. For example, a question asking "What is the total discount?" when the document shows original price and discounted price but not the discount amount requires arithmetic reasoning. A question asking "Is this contract favorable to the buyer?" requires legal knowledge to interpret contractual terms. These errors reflect fundamental limitations of the current framework—it excels at extracting and relating information that is explicitly present in the document but struggles with inference and domain-specific reasoning.

Model errors—cases where the model simply makes mistakes despite having correct inputs—account for 31-55% of errors depending on the task. On FUNSD, these include incorrectly linking a key to a value when multiple candidate values are nearby, or failing to recognize that a field is a key when its visual formatting is ambiguous. On SROIE, model errors often involve confusing similar entity types (item name versus store name). On RVL-CDIP, errors include misclassifying visually similar document types (scientific articles versus technical reports). These errors suggest that while the model has learned strong general representations, it still lacks some of the fine-grained discriminative ability that humans apply in ambiguous cases.

3.4 Cross-Task Transfer and Domain Adaptation

To evaluate the framework's ability to transfer across tasks and domains, we conducted two experiments: cross-task fine-tuning (pretraining on one task, then fine-tuning on another) and domain shift evaluation (training on one document domain, testing on another).

For cross-task transfer, we compared the unified framework pretrained on general documents to variants pretrained on task-specific data. Specifically, we pretrained separate models on (1) forms only (FUNSD-like documents), (2) receipts only (SROIE-like documents), and (3) mixed documents (our standard pretraining corpus). We then fine-tuned each on FUNSD and measured performance.

Table 6. Attention–Ground Truth Alignment.

Pretraining Data	**FUNSD F1**	**Δ vs Mixed**
Forms only	87.9 ± 0.6	-0.8
Receipts only	85.2 ± 0.9	-3.5
Mixed documents	88.7 ± 0.4	—
No pretraining	79.1 ± 1.2	-9.6

The results show that pretraining on mixed documents provides the best transfer, outperforming even pretraining on the target domain (forms only). This surprising finding suggests that exposure to diverse document types during pretraining helps the model learn more robust and generalizable representations than specializing on a single domain. The model pretrained on receipts only performs substantially worse (-3.5%), indicating that domain-specific pretraining can hurt transfer when the target domain differs significantly.

For domain adaptation, we trained on FUNSD and tested on CORD (a form understanding dataset from a different domain—business forms rather than government forms). Without any fine-tuning on CORD, the unified framework achieves 78.3% F1, compared to 84.2% when trained on CORD directly—a 5.9% gap. This zero-shot transfer performance exceeds LayoutLMv3's zero-shot performance (73.1% F1) by 5.2%, suggesting that the unified framework's explicit structural reasoning transfers better across domains than pure attention-based models. After fine-tuning on just 10% of CORD training data, the unified framework reaches 82.7% F1, recovering most of the performance gap and demonstrating strong few-shot adaptation.

3.5 Computational Efficiency Analysis

Table 5 compares the computational requirements of the unified framework against baseline models.

Table 5: Computational Efficiency Comparison

Model	Parameters	Training Time	Inference Speed	Peak Memory
LayoutLMv3	125M	6.2 days	42 docs/sec	28 GB
DocFormerv2	142M	7.8 days	31 docs/sec	32 GB
FormNetV2	118M	5.9 days	48 docs/sec	26 GB
Unified Framework	137M	7.0 days	38 docs/sec	30 GB
LayoutLMv3	125M	6.2 days	42 docs/sec	28 GB

Pretraining time on 64 V100 GPUs
**Inference throughput on single V100 GPU, batch size 1, FUNSD documents*
***Peak GPU memory during training, batch size 16*

The unified framework's parameter count (137M) falls in the middle of the baseline range, with slightly more parameters than LayoutLMv3 due to the additional graph module but fewer than DocFormerv2's encoder-decoder architecture. Pretraining time (7.0 days) is comparable to baselines, with the graph refinement adding approximately 15% overhead compared to LayoutLMv3. Inference speed (38 docs/sec) is faster than DocFormerv2 but slower than FormNetV2, primarily due to the cost of symmetric cross-attention over the full sequence. However, this speed is sufficient for most production use cases—processing 38 documents per second on a single GPU enables handling millions of documents per day with modest hardware.

We also measured the contribution of each component to computational cost. The graph module accounts for approximately 12% of total FLOPs (graph construction 3%, GNN propagation 4%, graph refinement 5%). Symmetric cross-attention adds approximately 8% overhead compared to asymmetric attention, as it requires computing attention in both directions. The visual encoder (ResNet-50) accounts for approximately 25% of total FLOPs, suggesting that using a more efficient visual encoder (MobileNet or EfficientNet) could substantially reduce computational cost with likely minimal impact on accuracy, since the transformer's cross-modal attention can compensate for less powerful visual features.

3.6 Interpretability Evaluation

To assess interpretability, we measured the alignment between model attention and human-annotated entity boundaries on FUNSD. For each predicted entity, we computed the IoU between the union of high-attention token bounding boxes (attention weight > 0.1 to the entity's first token) and the ground-truth entity bounding box.

Table 6: Attention-Ground Truth Alignment (Mean IoU)

Model	Entity Boundary IoU
LayoutLMv3	0.67 ± 0.18
Unified Framework	0.74 ± 0.15
Unified w/ Graph Visualization	0.81 ± 0.12

The unified framework's attention patterns align more closely with ground-truth boundaries (IoU 0.74) than LayoutLMv3 (IoU 0.67), suggesting that its predictions are more interpretable. When we include graph edge visualization (showing which tokens are connected by high-confidence edges), alignment improves further (IoU 0.81), as the graph makes structural relationships explicit.

In the user study with domain experts, participants rated the usefulness of attention visualizations at 3.8/5.0 for LayoutLMv3 and 4.3/5.0 for the unified framework ($p = 0.04$, Wilcoxon signed-rank test). Graph edge visualizations received an average rating of 4.6/5.0, with several participants noting that seeing explicit connections between keys and values helped them quickly verify predictions. Time-to-verification decreased from 18.3 seconds per document without interpretability outputs to 12.7 seconds with attention visualizations (30% reduction) and 9.4 seconds with both attention and graph visualizations (49% reduction), demonstrating that interpretability features provide practical value for human-in-the-loop workflows.

Qualitative analysis of attention patterns revealed that the unified framework learns meaningful structural attention: when processing a form field, the model attends strongly to nearby labels (keys), to other instances of the same entity type (enabling consistency), and to visual formatting cues (boxes, lines). Graph edges consistently connect semantically related elements—keys to their values, table headers to cells, section headers to contained text—with high-confidence edges (score > 0.8) achieving 87% precision against ground-truth relationships.

4. Discussion

The experimental evidence presented herein reveals that the unified multimodal framework has been a major factor in elevating document understanding across various domains to a level that is beyond that of recently published state-of-the-art models, both in text-centric and visually complex scenarios. It underscores the significance of interweaving the text, visual features, and layout structure of the same architecture instead of handling the modalities separately. The excellent results on FUNSD and SROIE indicate that the model is capable of capturing detailed spatial relationships that are vital for key–value extraction, while the improvements on DocVQA are indicative of better grounding of the textual content in the complex visual layouts. In the same way, the enhancements on RVL-CDIP and PubLayNet are evidence that the framework is capable of generalizing large-scale document classification and structural layout segmentation efficiently.

The pivotal feature of the method is the interplay between the symmetric cross-modal attention and the adaptive graph reasoning. Ablation results show that each mechanism alone cannot deliver the highest performance that is consistent; it is rather the combination that empowers the model to build more profound representations that mirror both semantic and structural dependencies. The attention bias that is aware of the spatial aspect facilitates the interaction between the text and the visual elements, whereas the refinement of the graph supports the understanding of the multi-column layouts, table structures, and hierarchical document regions. The synergy is a must for complicated documents where the semantic meaning comes not only from the content but from the arrangement as well..

Though it has been able to deliver a strong performance, a number of challenges still remain. For example, OCR errors still hamper the accuracy of form and receipt datasets, and semantic reasoning is still a problem in tasks that require implicit inference or domain-specific knowledge, like in DocVQA. Besides, the visually ambiguous document classes in RVL-CDIP, where the minor stylistic differences are not always captured by the patch-level representations, also perplex the model. The mentioned constraints point to the fact that forthcoming research should consider OCR-free pipelines, domain-aware reasoning modules, and hierarchical visual encoders that can identify the fine-grained stylistic cues.

The interpretability analysis reveals that the suggested framework helps the model to better align its attention with the regions identified by humans and at the same time, reduce the checking time for domain experts. This, in turn, means a lot of value for the healthcare, legal, and public sectors where transparency is strictly required and the stakes are high. The results point to the fact that multimodal interpretability instruments, primarily when integrating graph-level visualization, can profoundly improve human-AI collaboration by making the decision pathways more accessible and lessening the cognitive load during the validation process.

In summary, the findings highlight the importance of unified multimodal modeling as a foundation for robust and interpretable document AI systems. The proposed framework, which solves the problem of fragmentation in existing approaches and bases decisions on both visual and structural cues, is a springboard for more dependable and versatile document intelligence applications. The next steps in research should be to open up such potentialities by means of multilingual pretraining, strategies for low-resource adaptation, model compression for real-time deployment, and downstream business process integration..

5. Conclusions

The paper presented a single multimodal framework that combines text, layout, and visual information by using a symmetric cross-modal attention and adaptive graph reasoning for better contextual document understanding. The experimental results on five benchmark datasets show that the proposed framework consistently outperforms the recent state-of-the-art methods, especially in those tasks that demand fine-grained structural alignment and rich multimodal integration. The ablation analyses indicate that each architectural component—multimodal embeddings, spatially informed attention, and graph refinement—being a critical factor for strong generalization across different document types.

Besides quantitative performance, the framework also helps interpretability by giving more precise visual and structural explanations of its predictions. Such a feature facilitates human control and thereby, trust in automated systems, particularly in those areas where explainability is a prerequisite for compliance, auditing, and decision verification. The analysis of computational efficiency also indicates that the framework is still viable for large-scale deployment despite the increment in its modeling capacity.

To sum up, the unified multimodal framework represents a robust, generalizable, and interpretable solution to document understanding. Its solid empirical findings and adaptable architecture make it a good fit for real-world scenarios in finance, healthcare, legal services, and governmental administration. The next research direction could be enhanced semantic reasoning, multilingual extensions, OCR-free pipelines, and lightweight variants for on-device or low-resource environments. This study's findings set the stage for the advanced multimodal document AI systems that can accurately, transparently, and flexibly comprehend complex document structures.

References

[1] Wang D, et al. DocLLM: A layout-aware generative language model for multimodal document understanding. arXiv. 2023. https://doi.org/10.48550/arxiv.2401.00908

[2] Cooney C, Heyburn R, Maddigan L, O'Cuinn M, Thompson C, Cavadas J. Unimodal and multimodal representation training for relation extraction. In: Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science; 2022. https://doi.org/10.48550/arXiv.2211.06168

[3] Pyo B, Wang PSP. Key–value pair identification from tables using multimodal learning. Int J Pattern Recognit Artif Intell. 2023. https://doi.org/10.1142/S0218001423520092

[4] Gu Z, et al. XYLayoutLM: Towards layout-aware multimodal networks for visually rich document understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. https://doi.org/10.1109/CVPR52688.2022.00454

[5] Nguyen L, Scialom T, Staiano J, Piwowarski B. Skim-Attention: Learning to focus via document layout. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2021.

[6] Yu Y, et al. StrucTexT v2: Masked visual–textual prediction for document image pre-training. In: Proceedings of the International Conference on Learning Representations; 2023. https://doi.org/10.48550/arXiv.2303.00289

[7] Shi Y, Kim M, Chae YN. Multi-scale cell-based layout representation for document understanding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2023. https://doi.org/10.1109/WACV56688.2023.00366

[8] Jiang Z, Wang B, Chen J, Nakashima Y. ReLayout: Towards real-world document understanding via layout-enhanced pre-training. arXiv. 2024. https://doi.org/10.48550/arxiv.2410.10471

[9] Wei S, Xu N. PARAGRAPH2GRAPH: A GNN-based framework for layout paragraph analysis. arXiv. 2023. https://doi.org/10.48550/arXiv.2304.11810

[10] Chiron G, Arrestier F, Awal AM. Are attention blocks better than BiLSTM for text recognition? In: Proceedings of the International Conference on Machine Learning Technologies; 2023. https://doi.org/10.1145/3589883.3589914

[11] Shi D, Tao C, Rao A, Yang Z, Yuan C, Wang J. CrossGET: Cross-guided ensemble of tokens for accelerating vision–language transformers. arXiv. 2023. https://doi.org/10.48550/arXiv.2305.17455

[12] Zhang L, Xu D, Arnab A, Torr PHS. Dynamic graph message passing networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. https://doi.org/10.1109/CVPR42600.2020.00378

[13] Zhang C, Zhao Y, Yuan C, Tu Y, Guo Y, Zhang Q. Rethinking the evaluation of pre-trained text-and-layout models from an entity-centric perspective. arXiv. 2024. https://doi.org/10.48550/arxiv.2402.02379

[14] Aggarwal M, Sarkar M, Gupta H, Krishnamurthy B. Multi-modal association-based grouping for form structure extraction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2020. https://doi.org/10.1109/WACV45572.2020.9093376

[15] Sun Z, et al. CODA: Coordinating the cerebrum and cerebellum for a dual-brain computer-use agent with decoupled reinforcement learning. arXiv. 2025. https://doi.org/10.48550/arxiv.2508.20096

[16] Ramesh K, Sitaram S, Choudhury M. Fairness in language models beyond English: Gaps and challenges. In: Findings of the Association for Computational Linguistics; 2023. https://doi.org/10.48550/arXiv.2302.12578

[17] Wang W, et al. mmLayout: Multi-grained multimodal transformer for document understanding. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022. https://doi.org/10.1145/3503161.3548406

Author Contributions

Conceptualization, G.M.M.; Methodology, G.M.M.; Software, G.M.M.; Validation, G.M.M.; Formal Analysis, G.M.M.; Investigation, G.M.M.; Resources, G.M.M.; Data Curation, G.M.M.; Writing—Original Draft Preparation, G.M.M.; Writing—Review and Editing, G.M.M.; Visualization, G.M.M.; Supervision, G.M.M.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

All datasets used in this study are publicly available. Model code and scripts will be released upon publication.

Acknowledgments

The author acknowledges the valuable support provided throughout the research process.

Conflicts of Interest

The author declares no conflict of interest.