Multimodal Document Parsing and Data Structuring: A Review of Features, Techniques, and Challenges

Gabriel Ihuoma Lilian
Department of Computer Science
University of Port Harcourt,
Nigeria

Laud Charles Ochei
Department of Computer Science
University of Port Harcourt,
Nigeria

Uche Lilian Oghenekaro
Department of Computer Science
University of Port Harcourt,
Nigeria

ABSTRACT
The exponential growth of digital documentation across sectors such as finance, healthcare, legal services, and academia has created an urgent need for intelligent systems capable of extracting and organizing information from unstructured and semi-structured documents. Document parsing, the process of converting raw documents into structured, machine-readable data, has emerged as a critical enabler for downstream applications including information retrieval, knowledge base construction, and retrieval-augmented generation (RAG) systems. This paper presents a structured narrative review of the current state of multimodal document parsing and data structuring, covering key methodologies from traditional rule-based approaches to modern transformer-based architectures and large vision-language models. Core components such as layout analysis, content extraction (including text, tables, and mathematical expressions), and multimodal data integration are examined in detail. The review traces the evolution from early rule-based systems through statistical methods and machine learning approaches to the current dominance of transformer-based architectures, with particular emphasis on models such as LayoutLM, DocFormer, and Donut, and situates the recent emergence of general-purpose multimodal large language models as a further architectural direction. Widely used datasets and evaluation metrics are consolidated into a structured comparative framework, and reported benchmark results are compiled and analyzed to give a detailed, evidence-based comparison of technique performance, efficiency, robustness, and generalization. Persistent challenges are identified, including domain generalization, computational intensity, handling of noisy inputs, and evaluation reproducibility. The detailed technical coverage is drawn primarily from English-language literature and benchmarks; cross-lingual and low-resource-language document parsing is noted as a scope limitation and a direction for future work. The paper concludes by outlining future research directions, emphasizing the importance of developing robust, scalable, and generalizable document intelligence systems that can operate effectively across diverse document types and real-world conditions.

General Terms
Document Analysis, Machine Learning, Information Extraction

Keywords
Document Parsing, Data Structuring, Multimodal Learning, Transformer Architectures, Layout Analysis, Information Extraction, Document Understanding

1. INTRODUCTION

1.1 Background and Motivation
The digitization of information has fundamentally transformed how organizations create, store, and exchange knowledge. Electronic documents have increasingly supplanted paper documents as the primary medium for information exchange across industries, significantly expanding the diversity and complexity of document types encountered in daily operations: from contracts and invoices to academic papers and medical records. Organizations generate vast quantities of documents that contain valuable information embedded within diverse formats and structures.

Document parsing, also known as document content extraction, has emerged as an essential tool for converting unstructured and semi-structured documents into structured, machine-readable data. The process involves recognizing and extracting elements such as text, equations, tables, and images from document inputs while preserving their structural relationships; the extracted content is then transformed into structured formats such as Markdown or JSON, enabling integration into modern workflows [1]. Its importance has been further amplified by recent achievements in large language models (LLMs) and multimodal large models (MLLMs): document parsing plays an indispensable role in both knowledge base construction and training data generation for these systems [2], and it provides a foundation for retrieval-augmented generation (RAG) in practical fields and the automated construction of electronic storage and retrieval libraries [3]. However, a substantial proportion of historical records, academic publications, and legal documents remain in scanned or image-based formats, posing considerable challenges for information extraction, document comprehension, and retrieval [2]. Manual processing of such documents is time-consuming, error-prone, and resource-intensive, motivating automated solutions capable of handling document heterogeneity, complex layouts, and multimodal content.

Traditional parsing approaches that rely solely on textual content have proven inadequate for the rich, mixed formats found in real-world documents: modern documents combine textual content with visual elements (images, charts, logos), spatial layout information (headers, paragraphs, tables, columns), and structural hierarchies (sections, subsections, lists) [1]. Text provides semantic content, layout provides structural context, and vision provides stylistic and graphical cues; joint modeling of these signals captures synergies unavailable to unimodal approaches and enables more accurate parsing of complex documents [1]. Figure 1 summarizes the resulting processing pipeline, and Figure 2 contrasts it with the traditional unimodal pipeline it replaces.

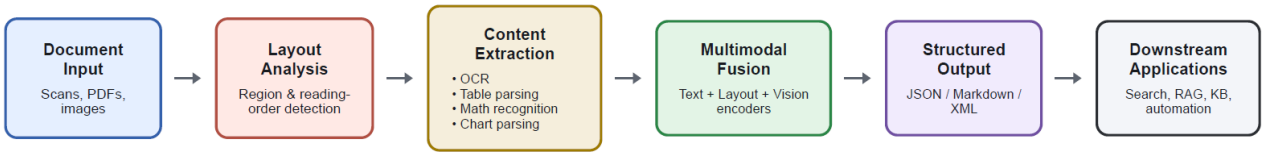

Figure 1. A multimodal document-processing pipeline: raw document input is processed through layout analysis and content extraction, fused across modalities, and converted into structured output for downstream applications.

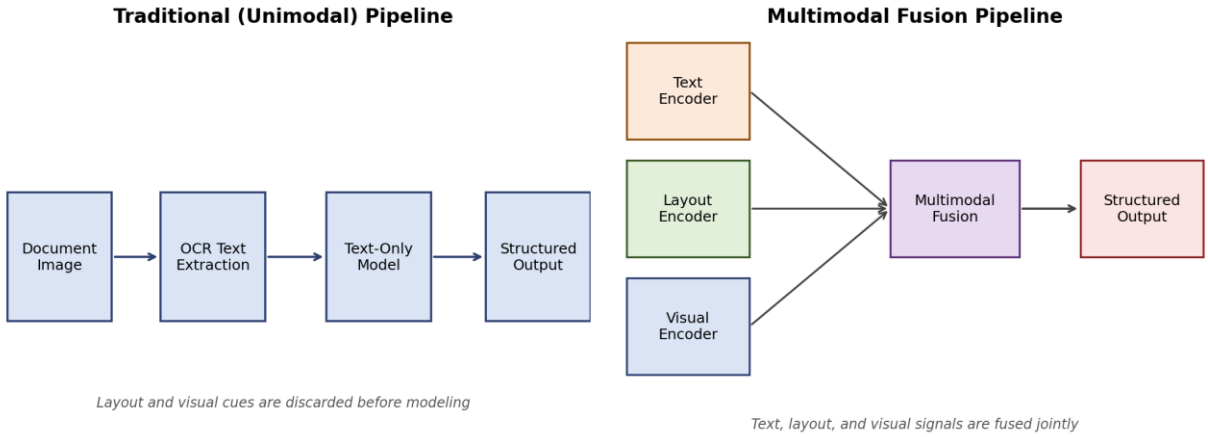

Figure 2. Comparison of modality interaction in traditional unimodal pipelines versus multimodal fusion pipelines. Traditional pipelines discard layout and visual cues once text is extracted, whereas multimodal architectures jointly encode text, layout, and visual signals before fusion.

1.2 Scope, Contributions, and Review Methodology

This paper presents a structured narrative review of features, techniques, and challenges in multimodal document parsing and data structuring. Its objectives are to: (i) provide a foundational understanding of core concepts, document types, and key tasks [1]; (ii) trace the evolution from rule-based and statistical methods to transformer-based architectures and large vision-language models [1, 2]; (iii) survey contemporary transformer-based models for multimodal document understanding and their architectural characteristics [4]; (iv) consolidate benchmark datasets and evaluation metrics into a single reference for comparative analysis [2]; (v) identify persistent challenges, including domain generalization, computational intensity, noisy inputs, and evaluation limitations [1]; and (vi) outline promising directions for future research [2].

The contribution of this review is a combined synthesis rather than a claim of first coverage: a historical narrative connecting rule-based, statistical, and deep-learning approaches to today's transformer-based methods; a five-part architecture-oriented taxonomy (layout-aware, vision-enhanced, structure-aware, OCR-free, graph-based) that organizes a fragmented literature into comparable categories; a consolidated mapping of models to tasks, modalities, datasets, and metrics; and a structured evaluation framework (Section 5) covering performance, efficiency, robustness, and generalization as explicit, separable comparison axes. Relative to [1] and [2], both valuable existing syntheses this review draws on, the contribution here is consolidating their coverage alongside the dataset and metric reference material and the evaluation framework into a single applied resource, rather than presenting new primary experimental findings — an approach consistent with structured review methodologies used in adjacent computer-vision domains such as video-based deepfake detection [5].

This review adopts a structured narrative approach rather than a systematic review: it borrows the transparency conventions of systematic reviews, including named databases, search terms, and explicit inclusion and exclusion criteria, while allowing purposeful selection of literature judged most relevant to a coherent, thematic account of a broad and rapidly evolving field; no claim of exhaustive or pre-registered coverage is made, since the diversity of architectures, datasets, and evaluation protocols in this area makes quantitative meta-analysis inappropriate. Literature was identified through Google Scholar, arXiv, IEEE Xplore, the ACL Anthology, the ACM Digital Library, and SpringerLink, using search terms including "document parsing", "document understanding", "visual document understanding", "layout-aware transformer", "OCR-free document understanding", and the names of major model families, covering January 2018 to July 2026 with seminal pre-2018 studies included for historical context. Papers were included if they proposed, evaluated, or reviewed document-parsing methods, presented novel architectures, datasets, or evaluation methodologies, and were published in peer-reviewed venues or widely recognized preprint repositories with sufficient technical detail for comparison; studies limited to conventional OCR without document understanding, generic vision tasks unrelated to document intelligence, or lacking sufficient technical detail were excluded. Additional studies were identified through citation snowballing from influential surveys, particularly [1] and [2]. Screened literature was organized thematically by historical development, architectural paradigm, input modality, understanding task, benchmark dataset, evaluation metric, and reported strengths and limitations to construct the comparative synthesis presented in the remainder of this paper.

2. RELATED CONCEPTS AND TERMINOLOGY

2.1 Document Parsing: Definition and

Subtasks

Document parsing is the process of converting unstructured and semi-structured documents into structured, machine-readable data. It involves recognizing and extracting elements such as text, equations, tables, and images while preserving their structural relationships, with extracted content transformed into formats such as Markdown or JSON to enable integration into downstream workflows [2]. It encompasses several subtasks that together support a comprehensive understanding of document content and structure: layout analysis [6], content extraction [7], logical structure recognition [7], entity extraction, and relationship extraction [1]. The resulting structured data, organized as JSON, XML, or relational tables, enables efficient storage, retrieval, and analysis of document content and supports downstream applications including search, automation, and machine learning [2]. Table I summarizes the key terminology used throughout this review.

Table I. Key Concepts and Terminology in Multimodal Document Processing

Term	Definition	Relevance
Document Parsing	Converting unstructured/semi-structured documents into structured, machine-readable data	Core task of this review
Data Structuring	Organizing extracted content into formats such as JSON, XML, or relational tables	Downstream enabler for search and automation
Layout Analysis	Identifying physical structures (text blocks, columns, tables, figures) and their spatial coordinates	Foundational subtask of document parsing
Multimodality	Joint use of textual, visual, and layout/spatial signals	Central theme of this review
Self-Attention	Mechanism allowing each token to attend to all others in a sequence	Basis of transformer architectures
OCR-Free Parsing	Direct image-to-structure parsing without a separate text-recognition step	Emerging architectural trend (e.g., Donut)
Graph-Based Reasoning	Representing documents as graphs of nodes (elements) and edges (relationships)	Addresses relational limitations of sequence models
Benchmark Dataset	Standardized, labeled corpus used to train and compare models	Enables reproducible evaluation

2.2 Document Types and Key Understanding Tasks

Documents vary in complexity based on their level of organization, formatting, and modality, and are generally categorized as structured, semi-structured, or unstructured [1, 2]. Several tasks recur across document-understanding systems. Optical character recognition (OCR) translates scanned or image-based text into editable text with bounding-box coordinates for each token, and serves as a foundational preprocessing step, though it still struggles with poor image quality, handwriting, and complex fonts [4, 8]. Layout analysis segments documents into logical zones such as paragraphs, tables, and figures, and determines reading order; research on the task began with rule-based and statistical methods in the 1990s and has, since 2015, been dominated by convolutional and transformer-based deep learning that treats layout analysis as pixel-wise segmentation [2, 6]. Information extraction and named entity recognition identify specific pieces of information such as names, dates, and key-value pairs; form understanding, a specialized case, is essentially a sequence-labeling task requiring models to comprehend entities across multiple modalities [1]. Table detection and parsing identifies tabular structures and extracts rows, columns, and cell values, a task that is vital in financial, scientific, and form-based documents, with end-to-end architectures such as TableNet and CascadeTabNet achieving strong performance on ICDAR benchmarks [2]. Document visual question answering (DocVQA) requires models to answer natural-language questions by jointly reasoning over a document's textual content and visual layout [9].

2.3 Modalities and the Transformer Foundation

Multimodal document understanding integrates several modalities to build a comprehensive interpretation of a document. The textual modality, processed with NLP techniques such as sentence segmentation and named entity recognition, forms the backbone of most tasks; models such as BERT and RoBERTa demonstrate strong contextual and semantic performance [10], building on the broader shift toward large pre-trained language models such as GPT-2 [11]. The layout modality captures the spatial organization of headers, paragraphs, tables, and images, and is essential in visually rich formats such as forms, invoices, and academic papers; layout-aware models such as LayoutLM leverage spatial embeddings for structural comprehension [4]. The visual modality covers scanned pages, diagrams, and embedded images, processed through OCR, object detection, and image classification; Vision Transformers and feature pyramid networks capture global context and multi-scale visual features [12]. The structural modality concerns how headings, subheadings, lists, and tables are organized hierarchically, supporting more precise extraction and downstream tasks such as summarization; graph neural networks and hierarchical transformers are commonly used to model such relationships [13].

Transformer-based models have become central to multimodal document understanding because of their capacity to model long-range dependencies and fuse heterogeneous modalities. The transformer architecture, introduced in [14], revolutionized sequence modeling by processing all tokens in parallel through self-attention: each position attends to all positions in the previous layer regardless of distance; multi-head attention runs several such attention mechanisms in parallel to learn different relational aspects; positional encodings inject sequence-order information since tokens are otherwise processed without inherent order; and, for multimodal tasks, cross-modality attention allows models to align and combine features from different modalities [14]. Figure 3 situates this architecture within the broader historical progression of document-understanding techniques.

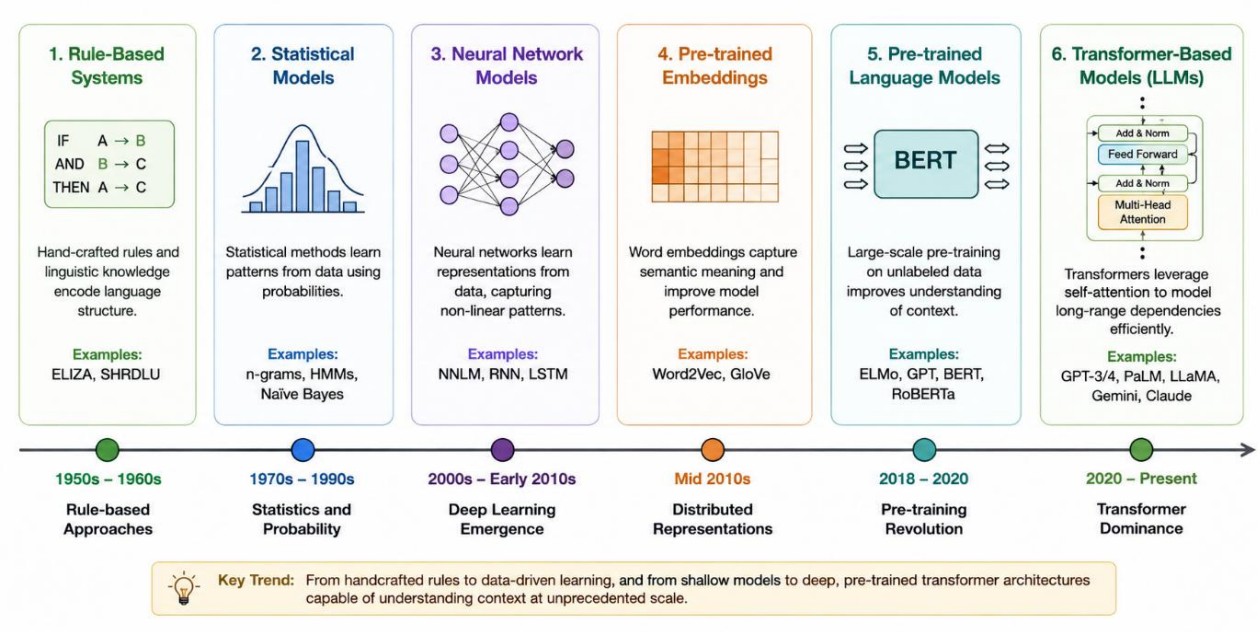

Figure 3. Timeline of evolution from rule-based systems to statistical models, neural networks, pre-trained embeddings, pre-trained language models, and transformer-based large language models.

3. REVIEW OF RELATED WORK

3.1 Evolution from Rule-Based to Deep Learning Approaches

Early document parsing relied on rule-based and heuristic systems using predefined rules and templates, with techniques such as top-down X-Y cuts and bottom-up clustering of connected components applied to layout analysis [2]. Such systems performed well on predictable, narrow document types but generalized poorly: minor layout deviations typically required substantial reconfiguration, and the approaches could not capture semantic or visual nuance [1]. Statistical methods, particularly Hidden Markov Models and Conditional Random Fields, later applied probabilistic modeling to document segmentation and information extraction, improving on rule-based systems but still depending on extensive hand-engineered features [1]. Classical machine learning methods, including decision trees, support vector machines, and random forests, subsequently offered greater adaptability by learning from labeled data, though they too required manual feature engineering and struggled with highly variable document structures [1, 2]. The advent of deep learning shifted the field toward data-driven, hierarchical representation learning: convolutional neural networks were applied to detect layout components such as tables and figures, while recurrent networks and BiLSTM-CRF architectures handled textual entity recognition [1, 2]. Early multimodal deep-learning models such as Chargrid combined character-level bounding boxes with 2D spatial position in a CNN-based architecture [1, 15], but models limited to a single modality generally failed to fully comprehend documents with complex visual structure or spatial dependency, motivating the multimodal transformer architectures reviewed next [2].

3.2 Transformer-Based Architectures for Multimodal Document Understanding

3.2.1 LayoutLM Series

LayoutLM [4] pioneered transformer-based multimodal document parsing by enhancing the BERT architecture with layout information such as the spatial position of text blocks, using a Faster R-CNN framework to extract regions of interest that are processed alongside text through masked visual-language modeling. Figure 4 illustrates the resulting architecture. LayoutLMv2 [16] extended the model with image embeddings and multimodal alignment pre-training, producing substantial gains on FUNSD, SROIE, and DocVQA, while LayoutLMv3 [17] further unified text and image masking and introduced word-patch alignment to improve cross-modal cohesion across both text-centric and layout-centric tasks.

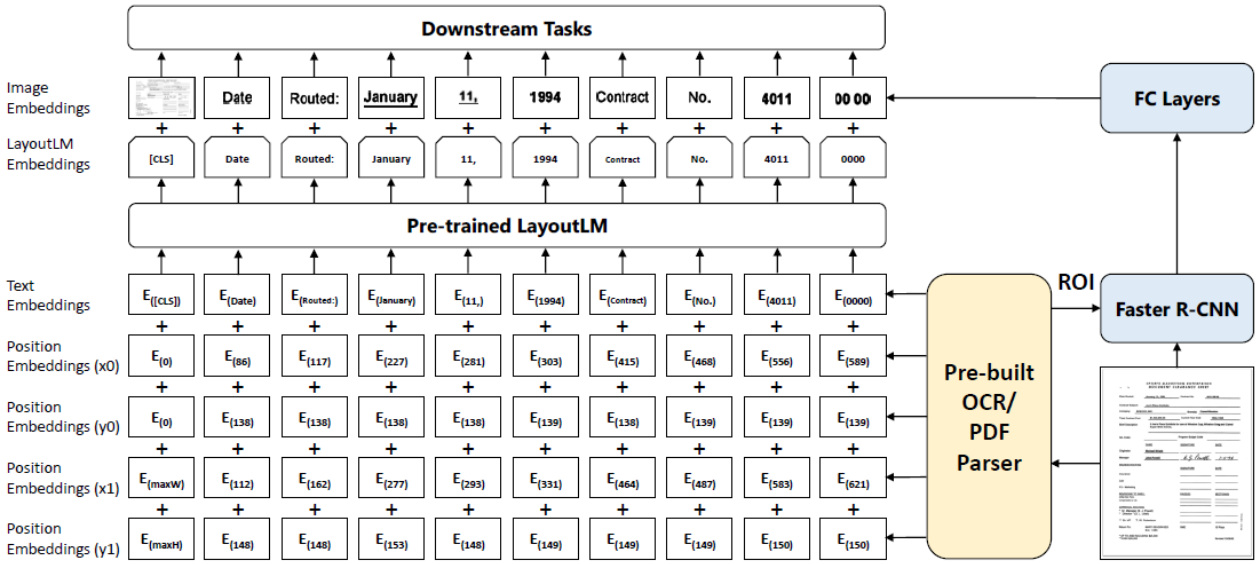

Figure 4. LayoutLM architecture combining text embeddings, layout (2D position) embeddings, and image embeddings within a pre-trained transformer, with visual regions of interest supplied by a Faster R-CNN detector [4].

3.2.2 DocFormer and Vision-Enhanced Models

DocFormer [18] integrates textual, layout, and visual features within a single joint encoder rather than a separate CNN-based visual extractor, encoding layout as a sequence of spatial tokens processed through multimodal self-attention; this fully transformer-based design avoids the feature-extraction bottleneck present in earlier LayoutLM variants [4].

3.2.3 Donut and OCR-Free Models

Donut [8] represents an OCR-free approach that processes document images directly with a Vision Transformer encoder and transformer decoder, eliminating cumulative OCR errors and demonstrating strong zero-shot and few-shot robustness to noisy scanned documents. Pix2Struct [19] similarly reframes document and UI understanding as image-to-text generation, pre-training on screenshot parsing rather than natural-image classification.

3.2.4 TILT and Related Encoder-Decoder Models

TILT [20] bridges text and image understanding through a unified T5-like [21] encoder-decoder that integrates layout as an attention bias and visual features via convolution, achieving strong performance on visual question answering and information extraction, particularly for documents such as academic papers where visual information is important [1]. A related OCR-free approach, Nougat [22], targets academic documents specifically, converting scanned scientific papers into markup that preserves mathematical notation, a capability the general-purpose OCR-free models above do not specifically optimize for.

3.2.5 FormNet and StructuralLM

FormNet [23] is tailored to form-document understanding, combining a graph-based transformer with 2D relational attention, Rich Attention over spatial relationships, and graph-convolution-derived Super-Tokens for neighbor-context awareness, achieving strong results on complex forms such as tax documents. StructuralLM [24] extends LayoutLM with cell- and segment-level structural representations and a cell-position classification pre-training objective, improving handling of hierarchical documents such as reports and contracts.

3.2.6 Emerging Architectures

DocLLM [25] incorporates layout-aware positional and structural embeddings into a large language model, improving summarization, extraction, and structured question answering over long documents. UDOP [26] offers a unified encoder-decoder that processes text, vision, and layout simultaneously without external OCR, learning directly from document pixels and achieving strong DocVQA and form-parsing performance [9]. The Document Image Transformer (DiT) [27] adapts Vision Transformers to document understanding through self-supervised pre-training on document images, avoiding token-based inputs and capturing structural and spatial features suited to visually complex or degraded documents [2].

3.2.7 Multimodal LLMs for Document Parsing

A parallel development to the purpose-built architectures above is the direct application of general-purpose multimodal large language models, including GPT-4V [28], Gemini [29], Qwen-VL [30], and InternVL [31], which accept document images directly and perform layout-aware reasoning, table interpretation, and structured extraction without a task-specific architecture. These models typically reduce OCR dependence further than even OCR-free specialist models and support flexible, prompt-driven structured output without task-specific fine-tuning, but they introduce challenges the specialist literature has already learned to manage: layout reasoning over dense multi-column or tabular regions remains less reliable than purpose-built spatial encoders [4, 23]; reproducibility is weaker, since outputs depend on proprietary, frequently updated models rather than a fixed, citable checkpoint; and hallucination, in the form of fabricated table values or invented text, is a documented failure mode largely absent from extraction-only architectures. This suggests the field is bifurcating rather than converging: specialized architectures remain preferable where reproducibility, auditability, and fixed deployment cost matter, such as regulated financial or legal processing, while general-purpose multimodal LLMs are increasingly favored for flexible, low-volume, or rapidly changing document types where fine-tuning a dedicated model is not cost-effective.

3.3 Graph-Based Approaches

Graph-based approaches explicitly model relationships between document elements using graph neural networks that

operate on documents represented as nodes and edges [13]. PICK [32] combines text and layout features with a graph-learning module that generates a soft adjacency matrix over text segments, decoded through BiLSTM and CRF sequence tagging, while Doc2Graph [33] frames document understanding more generally as a graph-construction problem over detected elements. GraphLayoutLM [34] explicitly models layout structure to capture spatial and hierarchical relationships for visually rich document understanding, and LayoutGCN [35] proposes a lightweight graph convolutional architecture that reduces computational load for real-time applications [1, 2].

3.4 Benchmark Datasets and Evaluation Metrics

Model evaluation in this field relies on a range of benchmark datasets spanning different document types and tasks, from handwriting and scene-text recognition (IAM Handwriting, SynthText, IIIT-5K) to layout analysis (PubLayNet [36], DocBank [37]), form and receipt understanding (FUNSD [38], CORD, SROIE), table parsing (TableBank, SciTSR), multimodal question answering and extraction (DocVQA [9], Kleister-NDA), document classification (RVL-CDIP [39]), and large-scale pre-training corpora (IIT-CDIP [40]). Table II consolidates these datasets by category, scale, and annotation type; Figure 5 illustrates, on a sample annotated page, the kind of token-level layout-category annotation that datasets such as DocBank provide (title, author, abstract, paragraph, figure, caption, table, and reference regions).

Table II. Benchmark Datasets for Document Parsing

Category	Dataset	Size	Annotation Type	Primary Use
OCR-Based	IAM Handwriting	1,539 pages	Text lines, words	Handwriting recognition
	SynthText	800k images	Word bounding boxes	Scene text detection
	IIIT-5K	5,000 word images	Word labels	Scene text recognition
Layout Analysis	PubLayNet	360k pages	Layout elements	Layout segmentation
	DocBank	500k pages	Token-level layout	Fine-grained parsing
Form / Key-Value	FUNSD	199 forms	Entities, relations	Form understanding
	CORD	1,000 receipts	Key-value fields	Receipt parsing
	SROIE	1,000 receipts	OCR, key-value	Invoice processing
Table Parsing	TableBank	417k tables	Table structures	Table detection
	SciTSR	15k tables	Table structures	Scientific tables
Multimodal	DocVQA	12k images	QA pairs	Visual question answering
	Kleister-NDA	3,300 contracts	Key fields	Contract extraction
Classification	RVL-CDIP	400k images	16 document classes	Document classification
Pre-training corpus	IIT-CDIP	11M+ images	Unlabeled scans	Large-scale pre-training

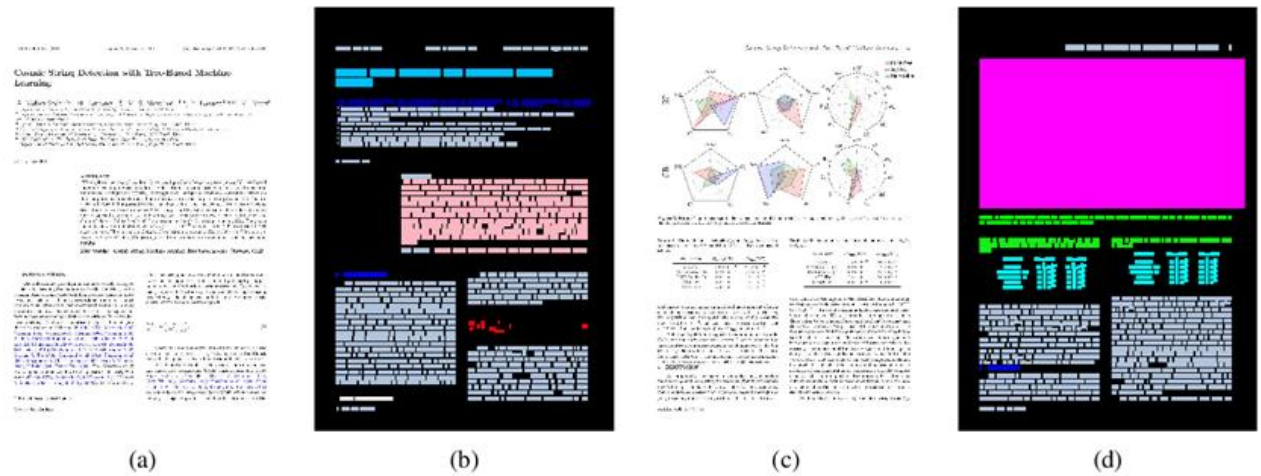

Figure 5. Example of token-level layout-annotation categories used by fine-grained layout datasets such as DocBank [37]: (a) a sample input document page; (b)-(d) the same page with colored regions denoting the semantic category assigned to each detected zone (title/author/abstract, body text, and table/figure regions).

Evaluation metrics vary with task objective: precision, recall, F1-score, and entity-level F1 for entity extraction and classification; intersection over union (IoU) and mean average precision (mAP) for layout detection and segmentation; accuracy, macro-F1, and micro-F1 for document classification; and ANLS (Average Normalized Levenshtein Similarity) for visual question answering. Table III summarizes these metrics alongside their formulas, target tasks, and trade-offs, adapted in part from the BLEU formulation of [41].

Table III. Evaluation Metrics for Document Parsing

Metric	Formula / Basis	Target Task	Strength	Weakness
Precision	TP/(TP+FP)	Entity extraction	Measures false positives	Ignores false negatives
Recall	TP/(TP+FN)	Entity extraction	Measures false negatives	Ignores false positives
F1-Score	2(P.R)/(P+R)	Entity extraction	Balances precision and recall	Assumes equal importance
Entity-level F1	F1 over full entity spans	Key information extraction	Reflects extraction usability	Token-level errors hidden
IoU	Intersection / Union	Layout detection	Spatial accuracy	Requires bounding boxes
mAP	Mean of per-class average precision	Layout / object detection	Standard detection benchmark	Sensitive to IoU threshold choice
Accuracy	Correct / Total	Document classification	Simple, interpretable	Misleading under class imbalance
Macro/Micro-F1	Per-class vs. pooled F1	Document classification	Class-balanced comparison	Macro-F1 sensitive to rare classes
ANLS	Normalized Levenshtein similarity	Visual question answering	Tolerant of minor OCR errors	Less strict than exact match
BLEU	Modified n-gram precision	Text generation	Fast, interpretable	Lacks semantic understanding

4. FINDINGS FROM THE REVIEW

4.1 Classification of Techniques by Features and Tasks

Contemporary document-parsing research reveals a clear taxonomy of methods categorized by the features they exploit: layout-aware models such as LayoutLM and its successors enrich textual embeddings with spatial coordinates to capture positional relationships critical for visually rich document understanding [1]; vision-enhanced models such as DocFormer and LayoutLMv2/v3 integrate visual features alongside text and layout to interpret graphics, logos, and handwriting; structure-aware models such as StructuralLM and FormNet incorporate structural semantics to enhance comprehension of hierarchical relationships [23, 24]; OCR-free models such as Donut and TILT bypass the need for separate OCR systems by processing document images directly [8]; and graph-based models such as PICK, GraphLayoutLM, and LayoutGCN employ graph neural networks to explicitly model relationships that sequence-based models may miss [13].

4.2 Comparative Strengths and Weaknesses

Table IV compares representative models across these categories on architectural paradigm, task specialization, and reported strengths and weaknesses. At the category level, the pattern is consistent: layout-aware models offer strong text-layout integration but remain dependent on OCR quality; vision-enhanced models add rich visual feature capture at a higher computational cost; structure-aware models capture hierarchical or relational structure well but generalize less widely outside their target document types; OCR-free models eliminate OCR error propagation but still struggle with dense tabular data; and graph-based models add explicit relational reasoning at the cost of graph-construction overhead. Figure 6 visualizes how these families' capabilities overlap: only a minority of systems jointly integrate text, layout, vision, and explicit relational reasoning, which remains an open architectural gap.

Table IV. Comparison of Document Understanding Models

Model	Core Architectural Paradigm	Primary Task Specialization	Strength	Weakness
LayoutLM	Multimodal BERT with layout/visual embeddings	Entity recognition, key-value extraction	Effective text-layout integration; high benchmark accuracy	Relies heavily on OCR quality; limited visual feature usage
LayoutLMv2	Two-stream multimodal transformer	Document layout and text understanding	Adds visual embeddings; more robust to complex layouts	Computationally expensive; still OCR-dependent
LayoutLMv3	Unified text and image masking	Document AI, multimodal understanding	Unified pre-training; improved cross-modal cohesion	High computational requirements
DocFormer	Joint multimodal encoder	Multimodal document understanding	Fully transformer-based; avoids feature-extraction bottlenecks	Requires large-scale training data; high memory use
Donut	OCR-free vision-encoder / text-decoder	OCR-free document parsing	Eliminates OCR dependency; robust to noisy documents	Struggles with highly structured tabular data
TILT	Generative encoder-decoder (T5-based)	Key information extraction	Strong text-image interaction; supports VQA	Limited to documents with strong text-image interplay
FormNet	Graph-based transformer	Form understanding	Captures relational dependencies; strong on forms	Limited generalization to non-form documents
StructuralLM	LayoutLM extension with structure-aware pre-training	Structured document parsing	Better hierarchical-organization capture	Less effective for unstructured or highly visual documents

Model	Core Architectural Paradigm	Primary Task Specialization	Strength	Weakness
PICK	Graph learning + transformer	Relationship extraction	Captures long-distance relationships	Graph construction adds preprocessing complexity
DocLLM	LLM with spatial encoding	Long-form document understanding	Leverages pre-trained LLMs; maintains spatial layout	Large model size; requires extensive fine-tuning
UDOP	Unified vision-text-layout encoder-decoder	DocVQA, form parsing, generation	Single architecture spans understanding and generation	Requires large multi-task pre-training corpus

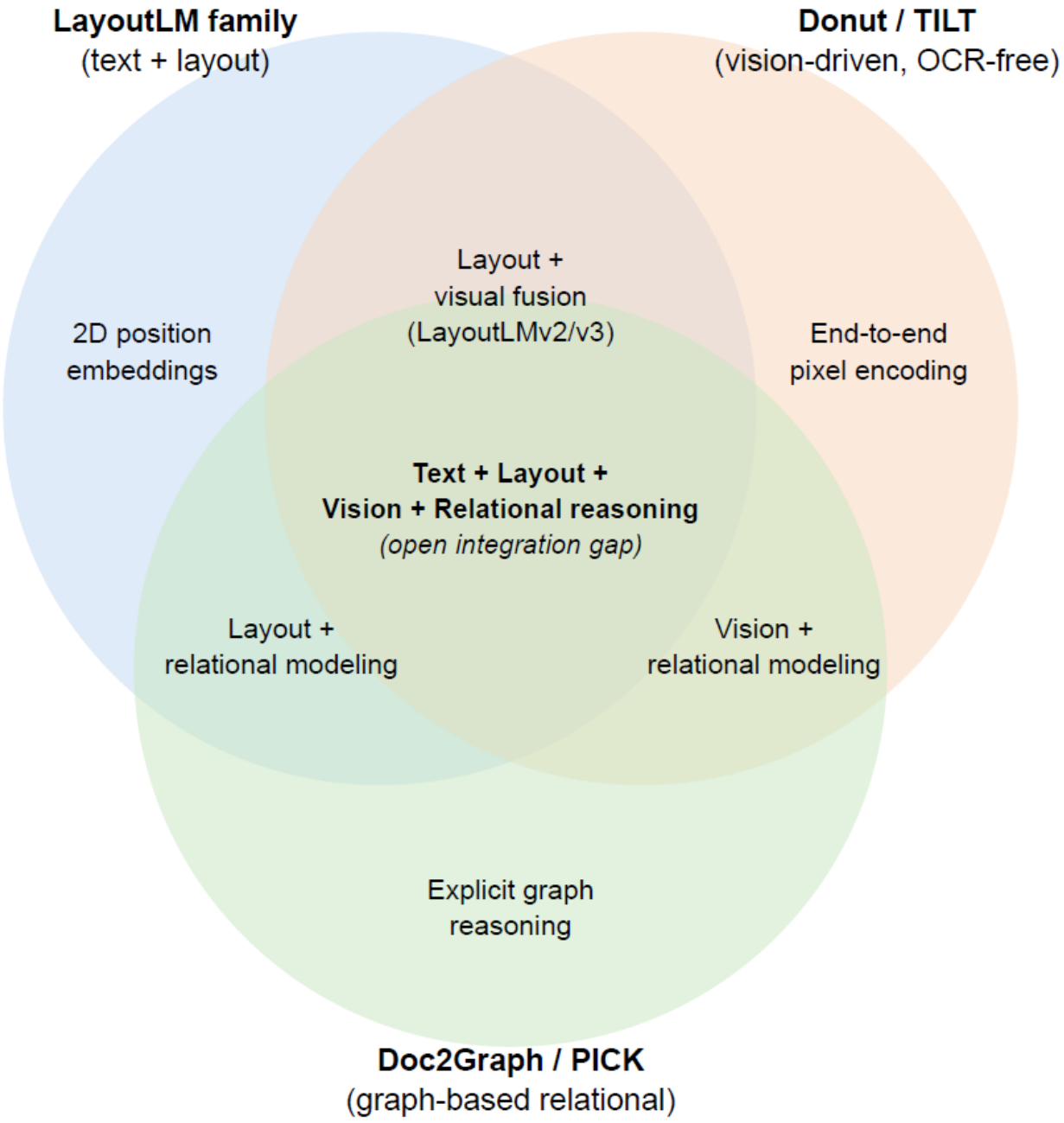

Figure 6. Overlapping capabilities across multimodal document parsing model families: layout-aware models (e.g., the LayoutLM series), vision-driven OCR-free models (e.g., Donut, TILT), and graph-based relational models (e.g., Doc2Graph, PICK).

4.3 Research Gaps, Limitations, and Emerging Trends

Several limitations recur across the reviewed literature. Models trained predominantly on benchmark datasets often underperform on unseen formats or domain-specific layouts [42]. Large-scale multimodal transformers demand high-end GPU resources, limiting deployment in cost-sensitive or low-resource environments (examined further in Section 6.2).

Existing models primarily rely on sequential tokenization and 2D positional embeddings, which struggle to capture relational dependencies between spatially separated but semantically related elements [4]. Many models rely on object detectors such as Faster R-CNN, designed for natural scene images rather than document layouts, for visual feature extraction [12]. Limited availability of diverse datasets constrains generalization across document variations [2], and real-world documents frequently present blurring, uneven formatting, or poor scan quality [1]. Evaluation practice itself reveals limitations, including benchmark overfitting, dataset bias, and incomplete reporting of implementation detail, examined in depth in Section 5.

The current research trajectory reveals several dominant trends: the rise of foundation models and prompting-based, zero-shot parsing; integrated language-vision pre-training; efficient models for edge deployment through pruning, quantization, and knowledge distillation that reduce computational footprint without significant accuracy loss [2]; and deeper multimodal data integration generally. Figure 7 depicts the structured evaluation framework, detailed in Section 5, that this review uses to organize the comparison of reported results across these four dimensions.

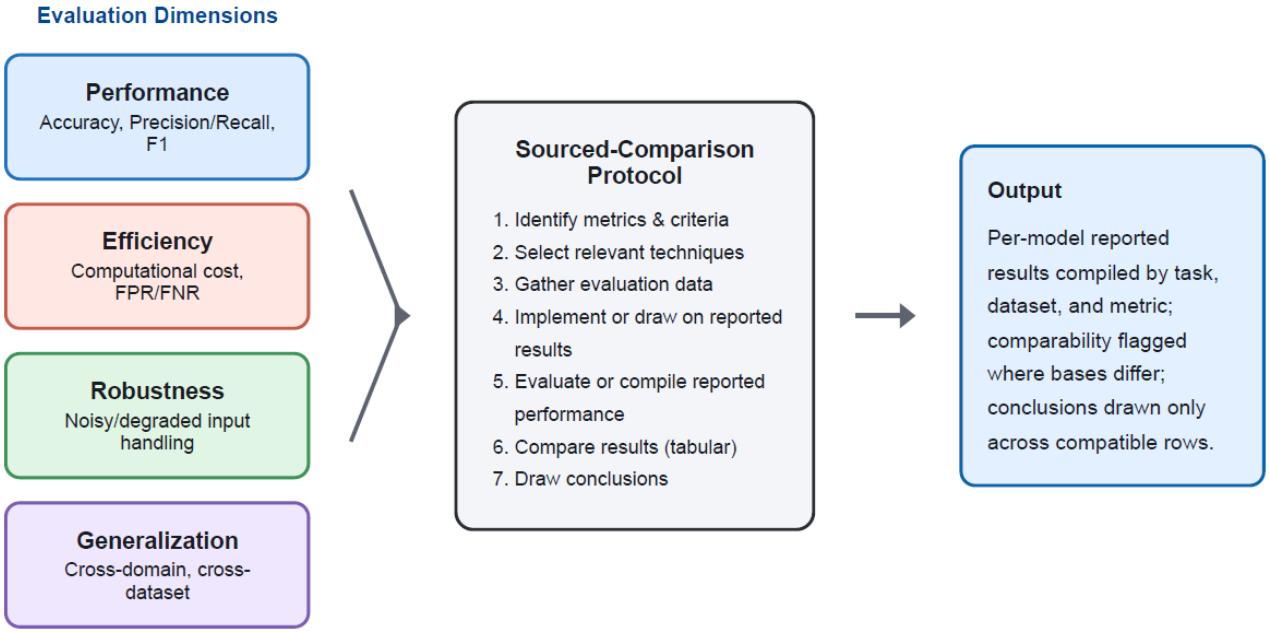

Figure 7. Structured evaluation framework for document parsing techniques: four evaluation dimensions (performance, efficiency, robustness, generalization) feed a sourced-comparison protocol whose output compiles each technique's own reported results rather than a single ranked score.

5. COMPARATIVE RESULTS AND EVALUATION

5.1 Structured Evaluation Framework

To compare document parsing techniques systematically, this review applies the seven-step protocol summarized in Figure 7: identify performance metrics and criteria (accuracy, precision and recall, F1-score, false-positive and false-negative rate, computational efficiency, robustness, generalization); select a representative set of techniques from the literature [2]; identify the evaluation data each reported result is drawn from; draw on each technique's own published implementation and reported results rather than re-implementing every method [1]; record the metric, dataset, and task precisely enough to assess comparability [2]; compare results in tabular form [1]; and draw conclusions only where the comparison basis is compatible. Section 5.2 applies this protocol to a subset of reviewed models using results reported in their original publications rather than a new implementation or evaluation performed by the authors of this review.

5.2 Reported Results Across Benchmarks

Table V compiles each model's own published result for a specific task, dataset, and metric combination. Because these models are frequently evaluated on different datasets, task formulations, or metrics, results are not normalized into a single ranked score; rows evaluated on incompatible bases are marked accordingly rather than forced into a misleading comparison. Figure 8 renders the directly comparable subset of Table V as bar charts: panel (a) shows the FUNSD entity-extraction F1 progression across LayoutLM, LayoutLMv2, and DocFormer, all reported on the same dataset with the same metric; panel (b) shows CORD document-parsing results for Donut and FormNet, annotated with the metric each figure represents, since Donut's reported accuracy and FormNet's reported F1 are not the same quantity.

Table V. Reported Results for Selected Document Parsing Models (Compiled from Original Publications)

Model	Task	Dataset	Metric	Reported Value	Source	Comparability Note
LayoutLM (base)	Entity / key-value extraction	FUNSD	F1	78.66	[4]	Baseline of this comparison set
LayoutLMv2 (base)	Entity / key-value extraction	FUNSD	F1	82.76	[16]	Directly comparable to row above

Model	Task	Dataset	Metric	Reported Value	Source	Comparability Note
DocFormer (base)	Entity / key-value extraction	FUNSD	F1	83.34	[18]	Directly comparable to rows above
Donut	Document parsing	CORD	Accuracy	91.10	[8]	Different metric from FormNet/CORD rows below
FormNet-A2	Key-value extraction	CORD	F1	97.10	[23]	Different metric from Donut/CORD row above
FormNet-A3	Key-value extraction	CORD	F1	97.28	[23]	Directly comparable to FormNet-A2

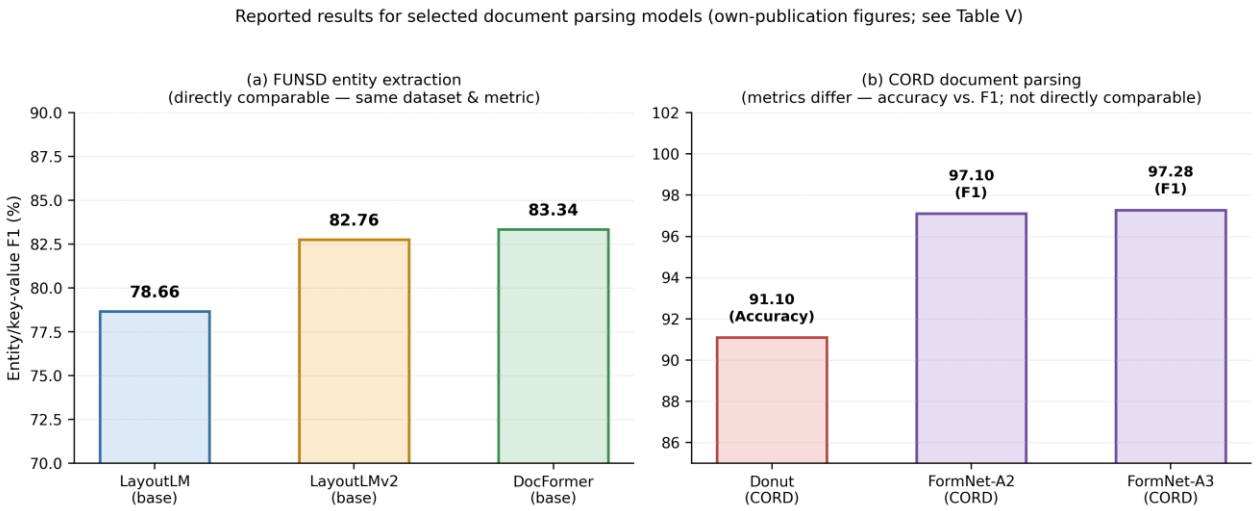

Figure 8. Reported results for selected document parsing models, compiled directly from their original publications (see Table V). Panel (a) compares FUNSD entity-extraction F1 on a common basis; panel (b) juxtaposes CORD results while flagging that the underlying metrics differ.

5.3 Analysis: What the Comparable Results Show

The FUNSD progression in panel (a) of Figure 8 is directly informative because all three figures share the same dataset and metric: LayoutLM reaches 78.66 F1 using text and layout alone; adding visual embeddings and multimodal alignment pre-training lifts LayoutLMv2 to 82.76 F1, a gain of 4.10 points; and replacing the CNN-based visual extractor with a fully transformer-based joint encoder lifts DocFormer to 83.34 F1, a further but visibly smaller gain of 0.58 points. Read alongside the architectural comparison in Table IV, this pattern is consistent with the qualitative trade-off reported throughout Section 3: the first move from text-only to text-plus-visual modeling yields the largest single improvement, while subsequent architectural refinement yields comparatively diminishing, though still positive, returns for added computational cost. The CORD comparison in panel (b) illustrates a different, cautionary point: FormNet's reported F1 (97.10-97.28) is numerically higher than Donut's reported accuracy (91.10), but the two figures answer different questions, an entity-level match rate versus a whole-document generation accuracy, so the gap cannot be read as FormNet outperforming Donut by nine points. This is precisely the kind of comparability failure that motivates the sourced-comparison protocol of Section 5.1 rather than a single leaderboard number, and it recurs throughout the wider literature: of the eleven models profiled in Table IV, the large majority have never been evaluated on a shared dataset and metric with more than one or two of the others, which is itself a finding about the state of evaluation practice in this field, not only a gap in this review's compilation (Section 6.2).

Beyond the directly reported numeric results, the three remaining evaluation dimensions from Figure 7, efficiency, robustness, and generalization, are assessed qualitatively in Table VI, drawing on the architectural characteristics and claims established throughout Section 3 rather than on new benchmarking. Layout-aware models carry the lowest inference-time compute cost but the highest OCR dependency, so their effective robustness is bounded by upstream OCR quality rather than by the document-understanding model itself. OCR-free models invert this trade-off, removing OCR dependency entirely and demonstrating stronger reported robustness to degraded scans [8], at the cost of weaker performance on dense tabular content where explicit cell structure matters. Graph-based models add a further processing stage, relationship graph construction, that increases compute cost modestly but improves robustness on documents where spatially distant elements are semantically linked. General-purpose multimodal LLMs sit apart from this trade-off space: their inference-time compute cost is the highest of any family considered, their robustness to noisy or unusual layouts is strong because they are not constrained to a fixed document-image training distribution, but this same flexibility introduces a distinct robustness failure mode, hallucinated content, that is largely absent from extraction-only architectures, and their reproducibility is comparatively low because outputs depend on proprietary, frequently updated models rather than a fixed, citable checkpoint. Taken together, Tables V and VI indicate

that no single family dominates across all four evaluation dimensions: the choice between them is properly read as a deployment decision governed by which dimension, accuracy, cost, robustness, or auditability, matters most for a given application, rather than as a search for a universally best technique.

Table VI. Qualitative Deployment and Robustness Profile by Model Family

Model Family	Relative Compute Cost	OCR Dependency	Robustness to Noisy Scans	Reproducibility / Auditability	Best-Suited Deployment Scenario
Layout-aware (LayoutLM family)	Low-Medium	High (external OCR required)	Bounded by OCR quality	High (fixed, citable checkpoints)	Clean, born-digital forms and invoices where a reliable OCR stage is already in place
Vision-enhanced (DocFormer, LayoutLMv2/v3)	Medium-High	Medium (partial visual compensation)	Moderate	High	Mixed-media documents (logos, stamps, photos) where visual cues carry semantic weight
Structure-aware (StructuralLM, FormNet)	Medium	Medium-High	Moderate; strong within form-like layouts	High	Highly hierarchical or tabular forms such as tax and government filings
OCR-free (Donut, TILT)	Medium	None (end-to-end pixels)	High for degraded scans; weaker on dense tables	High	Degraded, low-quality, or handwritten scans where OCR error propagation is the dominant risk
Graph-based (PICK, Doc2Graph, GraphLayoutLM)	Medium (added graph construction)	Medium-High	Moderate; benefits from explicit relations	Medium-High	Documents with spatially distant but semantically linked fields, such as multi-page contracts
General-purpose multimodal LLMs (GPT-4V, Gemini, Qwen-VL)	High (inference-time)	Low	High, but with hallucination risk	Low (proprietary, frequently updated)	Low-volume, rapidly changing document types where training a dedicated model is not cost-effective

The deployment-scenario column of Table VI is deliberately qualitative rather than numeric, consistent with the sourced-comparison protocol of Section 5.1: it reflects architectural characteristics and claims documented in the primary literature rather than new benchmarking performed for this review, and it is intended to guide practitioner selection rather than to rank families on a single axis. Together, Tables IV, V, and VI, read alongside Figure 8, give the fuller, more detailed evaluation picture that motivated this revision: architectural comparison, directly reported quantitative results where a shared basis exists, and a qualitative deployment profile across the dimensions for which no single shared numeric benchmark is yet available. No single family of techniques dominates every axis simultaneously: layout-aware models remain the most auditable and least computationally demanding but the most OCR-dependent; vision-enhanced and structure-aware models trade additional compute for richer representations of complex layouts; OCR-free models remove a brittle upstream dependency at the cost of weaker dense-table handling; graph-based models add explicit relational reasoning at modest extra construction cost; and general-purpose multimodal LLMs offer the broadest flexibility at the highest inference cost and the lowest reproducibility. This deployment-oriented reading of the evidence, rather than a single leaderboard ranking, is the practical takeaway of this comparative evaluation for practitioners selecting among these families for a specific document-processing task: the choice of model family is properly treated as a constrained optimization over accuracy, compute budget, OCR availability, and auditability requirements, rather than as a search for a single universally superior architecture. Section 6 builds directly on this evidence base: Section 6.1 draws out the implications of the comparative results above for both research and industrial practice, Section 6.2 examines the operational challenges that arise when deploying these techniques on real-world, noisy document collections rather than on curated benchmarks, and Section 6.3 outlines concrete directions through which the persistent limitations identified throughout Sections 4 and 5 can be addressed in future work.

6. DISCUSSION

6.1 Implications for Research and Practice

The evolution from rule-based parsing to transformer-based multimodal architectures has reshaped both academic inquiry and industrial application. These models, which integrate textual, visual, and spatial modalities, have delivered substantial improvements in tasks such as document question answering, form understanding, and table extraction [43]. This progression signals a shift toward holistic document understanding, in which systems interpret semantic content and structural relationships simultaneously; the integration of modalities requires fusion strategies such as early, late, and intermediate fusion, and the choice of strategy significantly influences the quality and coherence of the resulting data representation [1]. In practice, document-intensive sectors such as finance, healthcare, and legal services can now automate labor-intensive extraction processes, reducing cost and human error; the rise of OCR-free architectures such as Donut points toward a potential shift away from brittle OCR pipelines [8], and the growth of multimodal pretraining opens the door to domain-specific foundation models tailored to specialized industries [2].

Parsing quality has direct, technical consequences for RAG pipeline quality rather than merely motivational relevance: OCR errors introduce token-level noise that degrades retrieval precision; reading-order mistakes on multi-column or complex pages can silently interleave unrelated content into retrieved chunks that are locally coherent but globally wrong; poor table parsing either loses row and column relationships entirely or

embeds them as unstructured text, making numerical or relational queries unreliable regardless of retriever quality; and weak chunking that splits extracted text without respecting recovered logical structure routinely severs a heading from its body text or a table from its caption, degrading both retrieval relevance and the generator's ability to ground its answer [3]. Practitioners in the sectors motivating this review typically choose among open-source toolkits (for example Docling [44] and MinerU [45]), cloud document-intelligence services, and specialized RAG-ingestion platforms; selection generally turns on deployment constraints such as infrastructure cost, per-page pricing, and customization need, rather than on architecture-level accuracy comparisons alone, since production tools frequently combine or wrap several of the reviewed architectures internally.

6.2 Challenges in Real-World Deployment

Despite state-of-the-art benchmark results, multimodal document parsing systems face substantial operational challenges. Noisy inputs, including skewed scans, artifacts, and handwritten text, continue to impair accuracy, particularly for models dependent on clean OCR output [1]. This review's detailed technical coverage is drawn primarily from English-language literature and benchmarks, consistent with the current concentration of published work; cross-lingual and low-resource-language document parsing, covering scripts such as Arabic, Indic languages, and Thai, is flagged explicitly as a limitation of scope and a direction for dedicated future work rather than as a gap in the underlying field. The scarcity of annotated multimodal data exacerbates performance gaps for niche domains and underrepresented languages [1], and the computational demands of large-scale transformers, whose self-attention scales quadratically with input length, present barriers to deployment on edge devices or on-premises systems with constrained resources [14]. Effective real-world implementation further requires that these models integrate smoothly into existing enterprise document-processing systems [2].

6.3 Opportunities for Enhancement

Several directions can address the challenges above. Synthetic data generation offers a viable route to bridging annotation gaps [1], while unified pre-training across heterogeneous document types is a promising complementary direction [42]. Efficient models for edge deployment, and the related idea of dynamic modality selection so that a model processes only the modalities relevant to a given task, would extend deployability to resource-constrained settings [1]. Advancing multimodal fusion through unified representation frameworks [1] and integrating knowledge graphs into multimodal architectures for improved semantic reasoning [2] both remain open research avenues, alongside community-driven initiatives that prioritize open dataset releases, standardized evaluation protocols, and publicly shared pre-trained checkpoints to address the reproducibility concerns identified in Section 5.3.

7. CONCLUSION AND FUTURE DIRECTIONS

This paper has presented a structured narrative review of features, techniques, and challenges in multimodal document parsing and data structuring, tracing the evolution from rule-based and statistical methods through machine learning to the current dominance of transformer-based architectures, large vision-language models, and general-purpose multimodal LLMs. Key findings include the clear architectural taxonomy that now organizes the field (layout-aware, vision-enhanced, structure-aware, OCR-free, and graph-based models); the comparative strengths and weaknesses that make no single family dominant across performance, efficiency, robustness, and generalization simultaneously; the persistent challenges of domain generalization, computational cost, and evaluation reproducibility; and the bifurcation between specialized, auditable architectures and flexible, general-purpose multimodal LLMs, which is likely to define the field's near-term trajectory rather than convergence toward a single dominant architecture.

Multimodal models are transforming document processing by introducing more human-like interpretive capability: beyond text recognition, they incorporate spatial relationships, formatting signals, and visual content, an essential advantage for complex, visually rich documents such as invoices, forms, legal contracts, and scholarly articles [42]. By leveraging multiple modalities, these models reduce dependence on traditional OCR pipelines, offer greater adaptability to diverse layouts and formats, and improve accuracy in information-extraction tasks, with large-scale pre-training supporting stronger generalization and fostering innovation in both industrial automation and academic research.

Future work should prioritize five directions. First, model optimization and efficiency: lightweight architectures for edge and mobile deployment remain an active need, with dynamic modality selection as a promising complementary direction [1]. Second, richer datasets and benchmarking: the creation of comprehensive, diverse, large-scale multimodal datasets remains essential to the field's progression [1]. Third, cross-lingual and low-resource generalization, addressing the limitations of current models outside high-resource languages [2]. Fourth, integration of structured knowledge, particularly knowledge graphs, into multimodal architectures for improved semantic reasoning and structural understanding [1]. Fifth, advancing evaluation practice itself through community-driven initiatives that prioritize open-source dataset releases, standardized evaluation protocols, and publicly shared pre-trained checkpoints [1], so that future comparative reviews can rely on a larger set of directly comparable reported results than was available for Table V of this review.

8. ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their constructive comments, which substantially improved the clarity, structure, and technical depth of this manuscript.

9. REFERENCES

[1] Q. Zhang, V. S.-J. Huang, B. Wang, J. Zhang, Z. Wang, H. Liang, S. Wang, M. Lin, C. He, and W. Zhang. 2024. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. arXiv:2410.21169.

[2] K. Nandi and S. Siva Sathya. 2026. Visual document understanding: A comparative review of modern methods. In Computer Vision and Image Processing. Springer Nature Switzerland.

[3] L. Mei, S. Mo, Z. Yang, and C. Chen. 2025. A survey of multimodal retrieval-augmented generation. arXiv:2504.08748.

[4] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In Proc. ACM SIGKDD, 1192-1200.

[5] D. L. T. Bale, L. C. Ochei, and C. Ugwu. 2024. Deepfake detection and classification of images from video: A

review of features, techniques, and challenges. International Journal of Intelligent Information Systems 13, 2, 20-28.

[6] F. Delgado et al. 2026. A survey on reading order, table of contents, and structure extraction in document analysis. IEEE Xplore.

[7] B. Smock, R. Pesala, and R. Abraham. 2022. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In Proc. IEEE/CVF CVPR, 4634-4642.

[8] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. 2022. OCR-free document understanding transformer. In Proc. ECCV.

[9] M. Mathew, D. Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In Proc. IEEE/CVF WACV.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL-HLT, 4171-4186.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Technical Report.

[12] A. Dosovitskiy et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In Proc. ICLR.

[13] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. 2018. Graph attention networks. In Proc. ICLR.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, vol. 30.

[15] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Hohne, and J. B. Faddoul. 2018. Chargrid: Towards understanding 2D documents. In Proc. EMNLP, 4459-4469.

[16] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In Proc. ACL-IJCNLP, 2579-2591.

[17] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In Proc. 30th ACM International Conference on Multimedia.

[18] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha. 2021. DocFormer: End-to-end transformer for document understanding. In Proc. IEEE/CVF ICCV, 973-983.

[19] K. Lee, M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova. 2023. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In Proc. ICML.

[20] R. Powalski, L. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Palka. 2021. Going full-TILT boogie on document understanding with text-image-layout transformer. arXiv:2102.09550.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21, 140, 1-67.

[22] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic. 2023. Nougat: Neural optical understanding for academic documents. arXiv:2308.13418.

[23] C.-Y. Lee, C.-L. Li, T. Dozat, V. Perot, G. Su, N. Hua, J. Ainslie, R. Wang, Y. Fujii, and T. Pfister. 2022. FormNet: Structural encoding beyond sequential modeling in form document information extraction. In Proc. 60th Annual Meeting of the ACL, vol. 1, 3735-3754.

[24] Y. Li et al. 2021. StructuralLM: Structural pre-training for form understanding. In Proc. ACL.

[25] D. Wang, N. Raman, M. Sibue, Z. Ma, P. Babkin, S. Kaur, Y. Pei, A. Nourbakhsh, and X. Liu. 2024. DocLLM: A layout-aware generative language model for multimodal document understanding. In Proc. 62nd Annual Meeting of the ACL, vol. 1.

[26] Z. Tang, Z. Yang, G. Wang, Y. Fang, Y. Liu, C. Zhu, M. Zeng, C. Zhang, and M. Bansal. 2022. Unifying vision, text, and layout for universal document processing. arXiv:2212.02623.

[27] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, and F. Wei. 2022. DiT: Self-supervised pre-training for document image transformer. arXiv:2203.02378.

[28] OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774.

[29] Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. Google DeepMind Technical Report.

[30] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv:2308.12966.

[31] Z. Chen et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proc. IEEE/CVF CVPR, 24185-24198.

[32] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao. 2020. PICK: Processing key information extraction from documents using improved graph learning-convolutional networks. In Proc. 25th ICPR, 4363-4370.

[33] Y. Liu et al. 2023. Doc2Graph: Graph-based document understanding. In Proc. EMNLP, 4567-4580.

[34] Y. Li et al. 2023. GraphLayoutLM: Graph-based layout understanding for visually-rich documents. In Proc. ACL.

[35] D. Shi, S. Liu, J. Du, and H. Zhu. 2023. LayoutGCN: A lightweight architecture for visually rich document understanding. In Proc. ICDAR, 149-165.

[36] X. Zhong, J. Tang, and Q. Ye. 2019. PubLayNet: A large-scale dataset for document layout analysis. In Proc. IEEE/CVF CVPR.

[37] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou. 2020. DocBank: A benchmark dataset for document layout analysis. In Proc. COLING, 1234-1245.

[38] G. Jaume, H. K. Ekenel, and J.-P. Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In Proc. ICDAR-OST.

[39] A. W. Harley, A. Ufkes, and K. G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In Proc. ICDAR.

[40] D. Lewis, G. Agam, S. Argamon, O. Frieder, D.

Grossman, and J. Heard. 2006. Building a test collection for complex document information processing. In Proc. ACM SIGIR, 665-666.

[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Proc. 40th Annual Meeting of the ACL, 311-318.

[42] P. Kumaresan and M. B. Mohamad Amir Hamzah. 2025. Transformer based architectures for invoice layout classification. Multimedia University.

[43] W. Ahmad et al. 2024. A survey of recent approaches to form understanding in scanned documents. Springer Professional.

[44] Docling Team. 2025. Docling: An efficient open-source toolkit for AI-driven document conversion. arXiv:2501.17887.

[45] OpenDataLab. 2024. MinerU: An open-source solution for precise document content extraction. arXiv:2409.18839.