A Survey on MLLM-based Visually Rich Document Understanding: Methods, Challenges, and Emerging Trends

Yihao Ding[1], Siwen Luo[1] *, Yue Dai[2], Yanbei Jiang[2], Zechuan Li[2], Qiang Sun[1], Geoffrey Martin[3], Wei Liu[1], Yifan Peng[3]
[1]The University of Western Australia, Crawley, Australia
[2]The University of Melbourne, Melbourne, Australia
[3]Weill Cornell Medicine, New York, USA
Correspondence: {yihaoding, siwen.luo}@uwa.edu.au

Abstract

Visually Rich Document Understanding (VRDU) has become a pivotal area of research, driven by the need to automatically interpret documents that contain intricate visual, textual, and structural elements. Recently, Multimodal Large Language Models (MLLMs) have demonstrated significant promise in this domain, including both OCR-based and OCR-free approaches for information extraction from document images. This survey reviews recent advances in MLLM-based VRDU, highlighting emerging trends and promising research directions with a focus on two key aspects: (1) techniques for representing and integrating textual, visual, and layout features; (2) training paradigms, including pretraining, instruction tuning, and training strategies. Moreover, we address challenges such as data scarcity, handling multi-page and multilingual documents, and integrating emerging trends such as Retrieval-Augmented Generation and agentic frameworks. Our analysis offers a roadmap for advancing MLLM-based VRDU toward more scalable, reliable, and adaptable systems.

1 Introduction

Visually-Rich Document Understanding (VRDU) lies at the intersection of vision and language, aiming to extract and understand information from documents with multiple data modalities and complex layouts (Park et al., 2019; Ding et al., 2023). With the rapid digitization of physical documents and the widespread use of structured and semi-structured digital documents, the development of robust, generalizable VRDU frameworks has attracted significant attention for automating information extraction, improving accessibility, and enhancing decision-making across diverse domains such as finance, healthcare, and education.

Early VRDU frameworks relied on manually crafted rules and domain-specific heuristics (Watanabe et al., 1995; Seki et al., 2007), which experienced a sudden performance drop on unseen documents across domains or with diverse layouts. Conventional deep learning approaches employed CNNs (Katti et al., 2018; Yang et al., 2017) and RNNs (Denk and Reisswig, 2019) to leverage visual or textual features, facilitating more informative representations. However, these methods typically do not effectively integrate the diverse modalities in documents, limiting their capacity to capture the rich semantic structure inherent in visually rich documents. With the success of pretraining techniques in language modeling, numerous VRDU models (Huang et al., 2022; Hong et al., 2022; Lyu et al., 2024) have been pretrained on large-scale scanned or PDF document datasets, enabling more effective fusion of visual, textual, and layout features for robust multimodal representation. However, their effectiveness is constrained by the scope and diversity of their pretraining data, often necessitating substantial fine-tuning to achieve cross-domain generalizability.

Recently, MLLMs (OpenAI, 2024; Liu et al., 2024b), trained on massive visual and linguistic datasets, have demonstrated powerful representational capabilities and extensive world knowledge, enabling a deeper understanding of text-dense images with diverse visual appearances and complex spatial layouts. By combining the superior text understanding of LLMs (Touvron et al., 2023) with visual encoders (Dosovitskiy et al., 2020) that capture image content and layout information, MLLM-based VRDU frameworks have demonstrated strong performance across diverse document question-answering and information-extraction tasks, and generalizability across domains without task-specific fine-tuning.

This paper provides a comprehensive survey

*Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 13319–13340
July 2-7, 2026 ©2026 Association for Computational Linguistics

Model	Venue	Tasks	Mod.	LLM Backbone	Vision Encoder	PT	IT	FT	Pages	Prompt In.
OCR-Dependent										
ICL-D3IE (2023)	ICCV	KIE	T, L	GPT-3	–	✗	✗	✗	SP	ICL+Layout
DocLLM (2024a)	ACL	KIE, QA, DC	T, L	Custom	–	✓	✓	✗	SP	T+B+Q
LAPDoc (2024)	ICDAR	KIE, QA	T, L	Multiple	–	✗	✗	✗	SP	Rule
LMDX (2024)	ACL	KIE	T, L	Gemini-pro	–	✗	✗	✗	SP	ICL+Layout
ProcTag (2025)	AAAI	QA	T, V, L	GPT-3.5	–	✗	✗	✓	SP	Rule+CoT
DocKD (2024)	EMNLP	KIE, QA, DC	T, L	Custom	–	✗	✗	✓	SP	Gen by VL
DoCo (2024)	CVPR	KIE, QA, DC	T, L	Multiple	LayoutLMv3	✓	✗	✓	SP	I+Q
InstructDoc (2024)	AAAI	KIE, QA	T, V, L	FlanT5	LayoutLMv3	✗	✓	✓	MP	I+Q
LayoutLLM (2024)	CVPR	KIE, QA	T, V, L	Vicuna-7B-v1.5	LayoutLMv3	✗	✓	✓	SP	I+Q+CoT
LLaVA-Read (2024c)	preprint	KIE, QA	T, V, L	Vicuna-1.5 13B	Multiple	✓	✓	✗	SP	I+Q
LayTextLLM (2024)	ACL	QA, KIE	T, L	Llama2-7B-base	–	✓	✗	✓	SP	T+B
DocLayLLM (2024)	CVPR	QA, KIE	T, V, L	Llama2-7B-chat	Pix2Struct-Large	✗	✓	✓	SP	I+Q+B
LayTokenLLM (2025b)	CVPR	QA	T, L	Multiple	–	✓	✗	✗	MP	I+Q+L
GPE (2025a)	ICLR	KIE, QA	T, L	Multiple	–	✗	✗	✓	SP	T+B+Q
MDocAgent (2025)	preprint	QA	T, V	Multiple	ColPali, ColQwen2	✗	✗	✗	MP	I+Q
PDF-WuKong (2025)	preprint	QA	T, V	BGE-M3	IXC2-VL-4KHD	✗	✗	✓	MP	I+Q
DocAssistant (2025)	EMNLP	QA	T, V	InternVL2-Chat-2B	InternVL2 ViT	✗	✗	✓	SP	I+Q
AlignVLM (2025)	Neurips	QA	T, V	LLaMA-3.2 (1B, 3B)	SigLIP-400M	✓	✓	✓	SP	I+Q
DocThinker (2025)	ICCV	QA, KIE	T, V	Qwen2.5-VL (3B, 7B)	Qwen2.5-VL ViT	✗	✗	✓	SP	I+Q
OCR-Free										
KOSMOS-2.5 (2023)	preprint	QA, KIE	V	Custom	mPLUG-Owl VE	✗	✓	✓	SP	I+Q
mPLUG-DocOwl (2023a)	preprint	QA	V	mPLUG-Owl	mPLUG-Owl VE	✗	✓	✗	SP	I+Q
UReader (2023b)	EMNLP	QA	V	mPLUG-Owl	mPLUG-Owl VE	✗	✓	✗	SP	I+Q
TGDoc (2023)	preprint	KIE, QA	V	Vicuna-7B	CLIP-ViT-L/14	✗	✓	✓	SP	I+Q+B
UniDoc (2023)	preprint	KIE, QA	V	Vicuna-7B	CLIP-ViT-L/14	✗	✓	✓	SP	I+Q+B
DocPedia (2024)	SCIS	KIE, QA	V	Vicuna-7B	Swin Trans.	✓	✗	✓	SP	I+Q
HRVDA (2024a)	CVPR	KIE, QA	V	LLama2-7B	Swin Trans.	✓	✓	✗	SP	I+Q
Vary (2024)	ECCV	QA, DocRead	V	Multiple	CLIP, ViTDet	✓	✗	✓	SP	I+Q
mPLUG-DocOwl1.5 (2024)	EMNLP	KIE, QA	V	mPLUG-Owl2	mPLUG-Owl2 VE	✗	✓	✓	SP	I+Q
HVFA (2024)	Neurips	QA, Cap.	V	Multi (BLIP-2, etc.)	ViT/L-14	✗	✓	✗	SP	I+Q
Texthawk (2024a)	preprint	QA	V	InternLM-XC	ViT	✗	✓	✓	SP	I+Q
Texthawk2 (2024b)	preprint	OCR, Grd, QA	V	Qwen2-7B-Instr	SigLIP-SO400M	✗	✓	✓	MP	I+Q+Task
TextMonkey (2024c)	preprint	KIE, QA	V	Qwen-VL	Vit-BigG	✗	✓	✗	SP	I+Q
Llavar (2024d)	preprint	QA	V	Vicuna-13B	CLIP-ViT-L/14	✗	✓	✓	SP	I+Q
TokenCorrCompressor (2024b)	preprint	QA, Cap.	V	LLaMA-2	CLIP-ViT/L14	✗	✗	✓	SP	I+Q
DocKylin (2024a)	AAAI	QA	V	Llama2-7B-chat	Donut-Swin	✗	✓	✓	SP	I+Q
Marten (2025b)	CVPR	QA	V	InterLM2	InternViT-300M	✗	✓	✓	SP	I+Q
PP-DocBee (2025)	preprint	QA	V	Qwen2-VL-2B	ViT	✗	✗	✓	SP	I+Q
mPLUG-DocOwl2 (2025)	ACL	KIE, QA	V	mPLUG-Owl2	ViT	✓	✗	✓	MP	I+Q
TokenFD (2025)	ICCV	QA, KIE	V	InternLM (2B, 8B)	ViT	✓	✓	✓	SP	I + Q

Table 1: Comparison of existing MLLM-based VRDU frameworks. Mod.: Input modality; KIE: Key Information Extraction; QA: Question Answering; DC: Document Classification; T: Text; L: Layout; V: Vision; MP: Multi-Page; SP: Single Page; I: Image; Q: Question; B: Bounding Box; CoT: Chain of Thought; Cap.: Captioning; Grd.: Grounding; Task: Task Information; VL: Vision-Language.

of recent developments in MLLM-based VRDU frameworks. Previous surveys have either focused on a broad analysis of the diverse capabilities of MLLMs (Caffagni et al., 2024) or examined techniques applied to specific document understanding tasks, such as document layout analysis (Binmakhashen and Mahmoud, 2019), question answering (Barboule et al., 2025), and relation extraction (Delaunay et al., 2023). A recent study provides (Ding et al., 2025b) an overview of deep learning-based frameworks for VRDU but lacks a systematic perspective on MLLM-based approaches. In contrast, this paper provides an analysis of the MLLM-based VRDU frameworks from the aspects of **Framework Architecture** that covers both OCR- and OCR-free models (Sec 2), **Multimodal Representation** (Sec 3), **Training Strategies** (Sec 4), and **Inference Prompt Setting** (Sec 6). We also include a detailed discussion of the challenges of VRDU and provide a critical analysis of the trend and future directions (Sec 7). Notably, this survey is limited to methods that leverage MLLMs for document-level understanding, excluding multi-document applications, non-LLM-based methods, and MLLMs without VRD-specific adaptations.

2 Framework Architecture

General MLLM for VRDU. Many closed- (Team et al., 2024) and open-source (Chen et al., 2024) general-domain MLLMs have been widely adopted for VRDU tasks and have demonstrated promising performance [1]. However, the text-dense, visually rich, and layout-sensitive nature of VRDs exposes fundamental limitations of general-domain MLLMs when applied to VRDU, including weak layout inductive bias, sensitivity to OCR noise, and hallucination on these knowledge-

[1]Refer to Appendix C for performance analysis.

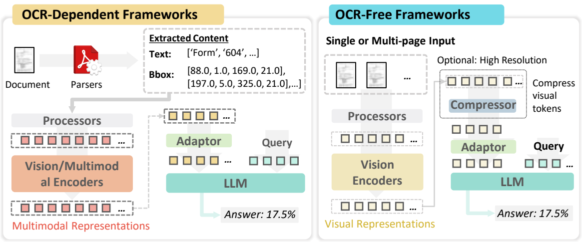

Figure 1: General OCR-dependent and OCR-free framework architectures.

intensive tasks. Moreover, the wide range of downstream VRDU applications necessitates specialized techniques that adapt existing LLM backbones (as shown in Figure 1) through VRDU-specific multimodal representations, training objectives, and inference paradigms. In addition, as VRDU tasks are often knowledge-intensive and safety-critical, locally tuning open-source general-domain LLMs on private document collections is essential for practical deployment in sensitive domains such as finance and industrial applications.

OCR-Dependent Frameworks. As shown in Figure 1, OCR-dependent frameworks leverage off-the-shelf tools to extract textual and layout information from scanned or PDF documents. This extracted data, in combination with the document image, is typically fed into multimodal encoders to generate joint representations. Some models (Wang et al., 2024a; He et al., 2023) input the extracted text directly into LLMs, while others (Luo et al., 2024; Zhu et al., 2025a) incorporate visual (Dosovitskiy et al., 2020) or multimodal encoders (Huang et al., 2022) to project those cues into language space via various adaptors or projects. These systems rely on external tools to capture structural information without extensive pretraining (e.g., text recognition). However, reliance on OCR or parsing tools can introduce cumulative errors, especially in handwritten or low-quality scanned documents, hindering the development of fully end-to-end models. Additionally, using low-resolution inputs may reduce the expressiveness of document representations, limiting the overall performance.

OCR-Free Frameworks. OCR-free approaches have been introduced for end-to-end VRD understanding tasks. These frameworks bypass text extraction by directly processing document images. Visual features are extracted via one or more vision encoders, fused with the user query, and decoded by an LLM to generate responses. Representative models include Donut (Kim et al., 2022), mPLUG-DocOwl (Ye et al., 2023a), and UReader (Ye et al., 2023b). Accurate comprehension of fine-grained text in these OCR-free settings requires high-resolution images, which, in turn, lead to lengthy visual sequences requiring visual compression modules (Liu et al., 2024a; Hu et al., 2025). Moreover, effective text recognition in these models often relies on large-scale pretraining or instruction-tuning to integrate textual and layout features via tasks such as text spotting (Liu et al., 2024c) and image captioning (Feng et al., 2024). This paradigm, however, demands substantial dataset construction and considerable computational resources, posing practical challenges.

3 Multimodal Representation

3.1 Text Modality

OCR-dependent methods rely on external tools to extract text for encoding, while OCR-free models use document images directly.

Text Encoding via LLM. Given the frequent text recognition challenges faced by MLLMs, stemming from low-resolution inputs or undertrained vision encoders, off-the-shelf OCR-extracted text is commonly embedded directly into LLM prompts to enhance document comprehension (Wang et al., 2024a; Kim et al., 2024) (see Figure 2). However, the extracted content is often unordered; to address this, frameworks such as ICL-D3IE (He et al., 2023) and LLaVA-Read (Zhang et al., 2024c) employ the XY-cut algorithm to reorder the text sequence. Additionally, to handle long documents, some methods segment the text into chunks, though this may introduce semantic discontinuities (Xie et al., 2025). In sum, directly adding extracted text to prompts improves context and reduces reliance on additional encoders; however, performance remains limited by OCR and LLM errors.

Text Encoding via Auxiliary Encoder. To enhance multimodal integration, many frameworks introduce auxiliary encoders to enhance text embeddings. Several methods (Luo et al., 2024; Zhu et al., 2025a) enhance text representation and multimodal fusion by feeding extracted text, image patches, and bounding boxes into pretrained LayoutLMv3 (Huang et al., 2022). Notably, Zhu et al. (2025a) propose a ROI Aggregation module

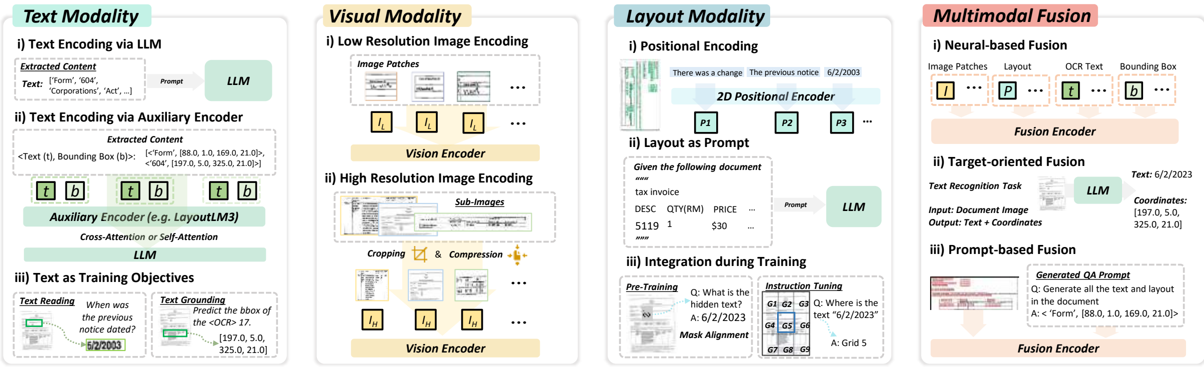

Figure 2: Multimodal feature representation and fusion mechanisms.

that aggregates fine-grained tokens (e.g., words) into object-level features (e.g., paragraphs), facilitating downstream object-level contrastive learning. Instruct-Doc (Tanaka et al., 2024) introduces an enhanced Q-Former (Li et al., 2023), termed *Document Former*, serving as a bridging module that integrates visual, textual, and layout information into the LLM input space via cross- and self-attention. In sum, external encoders improve representations but require additional pretraining and fine-tuning to align with LLMs' latent spaces.

Text as Training Objectives. Some frameworks rely exclusively on document images as input to predict answers. Models such as mPLUG-DocOwl (Ye et al., 2023a) and LLaVA-R (Zhang et al., 2024d), built upon mPLUG-Owl (Ye et al., 2023c), demonstrate strong OCR capabilities and are further instruction-tuned on diverse VRDU benchmarks. Other approaches incorporate text recognition, detection, and spotting tasks (Wang et al., 2023; Feng et al., 2023) to integrate text information. To better understand the hierarchical structure of documents, Hu et al. (2024, 2025) propose a multi-grained text localization task spanning the word-to-block level. While these methods deliver robust results using only visual inputs, they place heavy demands on pretraining and fine-tuning. Additionally, high-resolution images are often necessary to accommodate extremely long visual sequences and to preserve fine-grained features (Liu et al., 2024a; Yu et al., 2024a).

3.2 Visual Modality

To integrate visual information, OCR-dependent frameworks use extracted text and coarse visual cues, thereby enabling the use of **lower-resolution** images. In contrast, OCR-free frameworks require direct text recognition, demanding fine-grained perception and **high-resolution** inputs. See the Appendix A.3 for input resolution details.

Low Resolution Image Encoding. Some frameworks directly feed image patches into pretrained vision encoders to obtain patch embeddings (Xie et al., 2025; Tanaka et al., 2024). Others (Han et al., 2025; Luo et al., 2024; Liao et al., 2024) employ pretrained VRDU models, i.e., LayoutLMv3 (Huang et al., 2022), to extract multimodal-enhanced visual embeddings. Due to the limitations of low-resolution inputs in capturing fine-grained details, recent works have adopted dual-encoder architectures that process both low- and medium-resolution images (Ye et al., 2023b; Zhang et al., 2024c), followed by visual feature compression techniques to manage the increased feature volume. While using low-resolution images offers a straightforward pathway to multimodal understanding, achieving effective alignment often requires additional pretraining and instruction tuning. Moreover, the absence of fine-grained visual detail often necessitates additional OCR tools to extract text for accurate VRD interpretation.

High Resolution Image Encoding. To capture fine-grained level information for end-to-end training and inference, many frameworks support high-resolution image input. For ViT-style (Dosovitskiy et al., 2020) pretrained vision encoders, Hu et al. (2024) splits high-resolution images into predefined sub-images. To handle images of various shapes, UReader (Ye et al., 2023b) introduces a *Shape-Adaptive Cropping Module* that adaptively divides images into fixed-size sub-images using grids of various shapes. However, the image cropping may disrupt semantic continuity across sub-images. To address this, Liu et al. (2024c) introduced a *Shifted Window Attention* to enhance cross-sub-images connection via self-attention. In short, high-resolution images support fine-grained information extraction, but efficiently processing the resulting large number of visual tokens remains challenging, requiring a balance between resource usage and the number of visual tokens.

Visual Feature Compression. Yu et al. (2024a,b) utilize Q-Former (Li et al., 2023), while Liu et al. (2024c) adopts the *Resampler* from Qwen-VL (Wang et al., 2024b) to reduce the number of visual tokens. Considering the layout-aware nature of VRDs, Hu et al. (2024) introduces a convolutional module that preserves layout by compressing horizontal features and reducing the number of tokens. It further enhances this with layout-aware cross-attention to handle multi-page input. Liu et al. (2024a) use a *Content Detector* to filter non-informative tokens by segmenting text-rich regions, while Zhang et al. (2024a) propose eliminating low-information areas and clustering and aggregating the remaining features.

3.3 Layout Modality

Unlike natural scene images, VRDs feature dense text and complex layout structures. Methods for encoding layout information can be categorized into positional encoding-based, prompt-based, and task-oriented approaches.

Positional Encoding. OCR-dependent models use OCR tools to extract textual and layout information, combining text embeddings with 2D positional encodings (Xu et al., 2020) to incorporate layout into LLMs (Han et al., 2025; Tanaka et al., 2024). However, these approaches require extra training for feature alignment. In contrast, Zhu et al. (2025a) assigns unique positional embeddings to attention heads based on multi-dimensional layout features without altering the model architecture or requiring further pretraining. Wang et al. (2024a) treats layout as a separate modality and introduces disentangled spatial attention for cross-modal interactions without visual encoders. Zhu et al. (2025b) addresses long-context inference limits by encoding layout as a single token sharing the position with its text. However, these methods implicitly integrate layout information and rely heavily on large-scale pretraining, resulting in high computational costs and reduced effectiveness for tasks that demand explicit layout understanding.

Layout as Prompt. To integrate explicit layout information, some frameworks include layout details in prompts alongside the user query and document content. He et al. (2023) introduces an in-context learning based approach to incorporate layout-aware demonstrations into bounding box representations. Lamott et al. (2024) and Perot et al. (2024) encode layout into text sequence through rule-based verbalization or quantized coordinate tokens. These methods enable layout-awareness without training. However, these methods increase input length, rely on LLMs to interpret layout as text, and overlook visual cues essential for encoding relative positional information.

Integrating During Training. OCR-free frameworks incorporate text by formulating recognition and detection tasks that also aid in understanding layout (Wang et al., 2023; Feng et al., 2023). To further enhance this, some models (Wang et al., 2025b; Zhang et al., 2024c) leverage layout-aware pretraining tasks (Section 4.1) and layout-specific instruction-tuning tasks, such as visual grounding (Liu et al., 2024a,c) and table reconstruction (Liao et al., 2024). However, these methods typically require large-scale datasets for pretraining or instruction tuning, leading to substantial computational costs and data bottlenecks.

3.4 Multimodal Fusion

We categorize multimodal fusion methods into four types: direct, neural-based, task-oriented, and prompt-based. Direct fusion relies on simple feature summation or concatenation with alignment training, while this survey primarily focuses on the latter three approaches.

Neural-based Fusion. The simplest multimodal feature encoding uses external document encoders such as LayoutLMv3 (Xu et al., 2021), which fuse multimodal features via self- or cross-attention and leverage pretraining knowledge. Wang et al. (2024a) stands out by employing a layout-aware transformer with disentangled attention over text and spatial layouts, enabling effective document understanding without requiring image encoders. In OCR-free frameworks, visual encoders extract visual cues, with adaptors like LoRA (Yu et al., 2024b) or linear projectors (Zhang et al., 2024d; Wang et al., 2023) mapping features into the language space. Masry et al. (2025) propose a method that maps visual features to a weighted textual embedding to reduce misalignment issues observed in previous approaches. These neural-based fusion methods benefit from dedicated encoders or modified architectures, but often require extensive pretraining or SFT and face challenges in scalability, computational overhead, and adaptability to diverse document layouts.

Target-oriented Fusion. Target-oriented strategies establish multimodal connections through supervised objectives that span the input-to-output space (Hu et al., 2024) and are widely applied to text and layout features in OCR-free frameworks. For instance, in text recognition tasks, models are trained to map visual features directly to text and spatial coordinates, thereby aligning fusion with task-specific goals. While these approaches improve end-to-end multimodal integration, they also increase demands on data preparation, annotation quality, and training complexity in practice.

Prompt-based Fusion. Prompts for multimodal tasks may include text, images, and bounding box coordinates. While many frameworks adopt Layout-as-Prompt strategies to encode layout information, others use Chain-of-Thought (CoT) reasoning to further enhance multimodal learning. For example, Luo et al. (2024) utilizes a *LayoutCoT* approach that divides reasoning into question analysis, region localization, and answer generation, explicitly modeling spatial layout. Liao et al. (2024) leverages CoT pretraining and CoT annealing to support layout-aware reasoning for VRDU. However, these methods often depend on predefined reasoning strategies, intermediate-step evaluations, and well-trained prior frameworks, limiting their generalizability.

4 Training Paradigms

To facilitate multimodal understanding, instruction following, and domain adaptation, various training tasks and strategies have been developed, as illustrated by Figure 3.

4.1 Pretraining Strategies

To enhance mono- and multi-modal document understanding, VRDU frameworks adopt various self-supervised pretraining tasks, such as masked information modeling and cross-modality alignment (Ding et al., 2025b). OCR-dependent frameworks typically utilize pretrained VRDU models or vision encoders to obtain enriched multimodal representations. Some models propose additional self-supervised learning tasks (e.g., Li et al. (2024) applies object-level contrastive learning between visual and multimodal features). Wang et al. (2024a) introduces a transformer architecture with disentangled spatial-text attention to perform block-wise text infilling to enhance text-layout correlation modeling. OCR-free frameworks (Zhang et al., 2024c; Hu et al., 2024) focus on pretraining tasks like text recognition, detection, and captioning to integrate text and layout information. Hu et al. (2025) further targets multi-page layout coherence. Feng et al. (2024) aligns frequency features with LLMs through text-centric pretraining. Although these self-supervised tasks are effective in fusing multimodal features and learning general knowledge, they remain computationally intensive and often lack instruction-based tuning, limiting their capacity to follow real-world user instructions.

4.2 Instruction Tuning

To benefit task orientation in LLM-based frameworks, many VRD approaches, following InstructGPT (Ouyang et al., 2022), are trained on instruction-response pairs to better align model outputs with user prompts. Pretraining tasks such as text reading, recognition, and image captioning are reformulated as instruction-based formats, with images paired with task descriptions. Beyond improving multimodal fusion, goal-oriented tasks, including VRD question answering (Ding et al., 2024b), key information extraction (Ding et al., 2023), and VRD classification (Harley et al., 2015), are conducted on large-scale datasets. For better generalizability, some frameworks synthetically generate large instruction-tuning datasets (See Appendix B for more details). To further improve localization and information extraction, Wang et al. (2023) and Feng et al. (2023) propose predicting answers alongside bounding boxes, thereby enhancing the framework's reliability. Instruction tuning not only strengthens user query understanding but also boosts multimodal fusion. Instruction tuning on large-scale datasets substantially enhances zero-shot performance. However, the requirement for extensive training data leads to substantial resource consumption. Furthermore, synthetic datasets, often generated with off-the-shelf OCR tools and LLMs, may yield low-quality QA pairs, particularly in low-resource domains such as scanned documents, thereby impacting zero-shot performance.

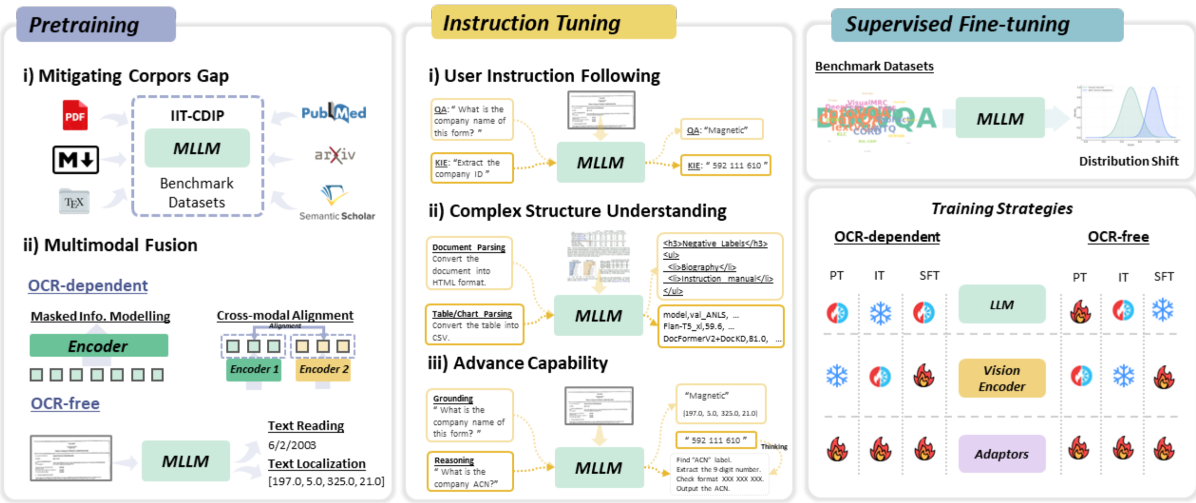

Figure 3: MLLM-based VRDU framework training paradigms.

4.3 Training Strategies

MLLM-based document understanding frameworks typically consist of multiple sub-modules to encode multimodal information and are trained in a stepwise manner. Few frameworks leverage in-context learning (He et al., 2023) or multimodal prompts (Perot et al., 2024) to develop training-free architectures. The majority, however, involve pretraining to capture general-domain knowledge, followed by instruction tuning to improve the interpretation of user prompts. Furthermore, some frameworks are subsequently **Supervised Fine-Tuned** on benchmark datasets (Wang et al., 2024a; Zhu et al., 2025a) or a synthetic set (Kim et al., 2024) to enhance domain-specific adaptation. To integrate multimodal information, these frameworks mainly employ an LLM with various multimodal encoders (Han et al., 2025; Xie et al., 2025), sometimes incorporating adaptors (Hu et al., 2024; Lu et al., 2024) or linear projectors (Park et al., 2024) for fusion or alignment. Depending on the training stage, sub-modules may be either trainable or frozen, balancing the acquisition of new knowledge with the preservation of valuable information from the original backbone.

LLM Backbone. As most LLMs are extensively pretrained on large-scale datasets and capture broad knowledge, many frameworks freeze the LLM, using it solely to generate human-understandable outputs. In frameworks involving pretraining or instruction tuning (Zhang et al., 2024a; Liu et al., 2024a), freezing the LLM backbone helps preserve its knowledge and reduce training costs. However, some approaches enable LLMs to be trained during continued pretraining (Zhu et al., 2025b) or instruction tuning (Liao et al., 2024) to better capture VRD domain knowledge and enhance multimodal alignment. In supervised fine-tuning stages, the LLM backbone is typically made trainable to adapt to the target domain (Zhang et al., 2024d).

Multimodal Encoders. They are employed to encode multimodal features, which are subsequently aligned with LLM text representations by projectors or adapters. Similar to LLM backbones, vision (Dosovitskiy et al., 2020), and multimodal encoders (Huang et al., 2022) are often kept frozen during pretraining to preserve learned knowledge (Yu et al., 2024b; Zhang et al., 2024d). Feng et al. (2024) use a Swin Transformer to encode frequency-domain images, pretrained from scratch. To enhance multimodal feature learning, Li et al. (2024) make the ViT encoder trainable while freezing LayoutLMv3, enabling knowledge distillation via contrastive learning. During instruction tuning, vision encoders are typically unfrozen to improve alignment and task-specific adaptation (Zhang et al., 2024a; Liu et al., 2024a). Conversely, in dual-encoder frameworks, vision encoders with inputs at diverse resolutions are often frozen to enhance the representation of hierarchical inputs. In supervised fine-tuning, there is no standard practice for encoder trainability.

Projectors and Adaptors. They play a crucial role in feature alignment and lightweight tuning. Projectors are typically employed to align visual or layout features with the LLM input space

(Park et al., 2024) and encode layout information (Tanaka et al., 2024). These modules are mainly trainable throughout the entire training process. Adaptors, on the other hand, are designed for efficient, task-specific tuning, often leveraging LoRA-style updates (Ye et al., 2023a; Hu et al., 2024) or cross-attention mechanisms (Liu et al., 2024c; Yu et al., 2024a) to integrate multi-aspect inputs with minimal parameter changes. Plug-and-play components, such as visual abstractors (Ye et al., 2023a) or compressors (Hu et al., 2025), have also been introduced to reduce the dimensionality of visual features. These adaptors are usually trained during instruction tuning or during supervised fine-tuning.

5 Datasets

5.1 Pretraining Datasets.

The goal of pretraining is to enhance multimodal understanding and improve generalization across VRDU tasks. MLLM-based approaches commonly perform continued pretraining on large-scale, cross-domain document collections such as IIT-CDIP (Lewis et al., 2006), which contains over 6 million scanned documents across diverse domains, though lacking explicit layout annotations, often supplemented with OCR-derived bounding boxes. RVL-CDIP (Harley et al., 2015), a curated subset with 400,000 documents across 16 categories, is widely used for document classification. Beyond these general-purpose datasets, recent frameworks (Zhang et al., 2024d; Wang et al., 2023) have introduced self-collected datasets to target domain-specific or task-oriented scenarios, including slide decks (Feng et al., 2024), academic papers (Wang et al., 2024a), and other structured document types (Yu et al., 2024b).

5.2 Instruction-tuning Datasets.

Instruction-tuning aims to enhance a model's understanding of user queries. Many frameworks (Zhang et al., 2024b; Park et al., 2024) perform instruction-tuning directly on benchmark document collections to improve downstream task performance. Others (Luo et al., 2024; Liu et al., 2024a) generate large-scale synthetic datasets using OCR tools to extract text and layout information from VRD-related benchmarks such as layout analysis (Zhong et al., 2019) and document classification (Harley et al., 2015). Instruction-response pairs are then created based on predefined task definitions. Some frameworks also construct their own multi-domain datasets to improve generalizability and prevent data leakage (Wei et al., 2024; Feng et al., 2023). Instruction-tuning is critical for domain adaptation and accurate instruction interpretation. As shown by Table 7, some frameworks increasingly generate synthetic instruction-tuning datasets tailored to their architectures, prioritizing alignment over generalizability achieved through benchmark-based tuning.

5.3 Benchmark Datasets

Key Information Extraction Benchmarks for Key Information Extraction (KIE) are shifting from early schema-constrained tasks (e.g., SROIE (Huang et al., 2019), FUNSD (Jaume et al., 2019)) toward larger, multilingual, cross-domain, multi-page, and open-vocabulary challenges. While form-like structures (e.g., DocILE (Šimsa et al., 2023), Form-NLU (Ding et al., 2023)) still dominate the landscape, modern resources such as KVP10k (Naparstek et al., 2024) and CC-OCR-KIE (Yang et al., 2024) focus on *open-category* extraction without predefined schemas. Furthermore, a clear trend of dataset consolidation and multilingual expansion has emerged.

Visual Question Answering. has undergone a comparable evolution, shifting from early single-page, text-centric retrieval to benchmarks that probe multiple dimensions of complexity. This progression is reflected in broader multilingual coverage (e.g., MTVQA (Tang et al., 2025), JDocQA (Onami et al., 2024)) and more diverse, multi-domain settings (e.g., DUDE (Van Landeghem et al., 2023)). Recent datasets increasingly emphasize long-context comprehension over multi-page documents: benchmarks such as LongDocURL (Deng et al., 2025), BRIDGE (Xiang et al., 2026) and MMLongBench-Doc (Wang et al., 2025a) contain documents averaging dozens of pages and often demand non-trivial cross-page evidence aggregation and reasoning. In parallel, reasoning requirements have deepened toward domain-specific expertise, as illustrated by vision-essential physics problem solving in SEEPHYS (Xiang et al., 2025). Finally, dataset scale has expanded substantially, reaching millions of instances in collections such as MMVQA (Ding et al., 2024a), thereby enabling rigorous stress-testing of the capacity and reasoning limits of modern multimodal models.

6 Inference Prompt Setting

MLLM-based frameworks adopt diverse prompt formats depending on their architecture. For OCR-free frameworks in Table 1, the prompt typically includes a document image, occasionally multiple pages (Hu et al., 2025; Wang et al., 2025b), alongside a textual user query. Some frameworks not only predict answers to user queries but also localize bounding boxes, often requiring an additional prompt for localization (Wang et al., 2023; Feng et al., 2023). OCR-dependent frameworks first preprocess input using off-the-shelf tools to extract textual and layout information. Vision-free models (He et al., 2023; Wang et al., 2024a) process only the extracted content alongside the query. In contrast, vision-dependent models also incorporate the document image into the vision (Xie et al., 2025) or into multimodal encoders (Liao et al., 2024), aligning visual and textual features for the final prediction. Furthermore, some frameworks integrate layout information into prompts via bounding boxes (Zhu et al., 2025a) or markdown-style formatting. The inference strategies are closely tied to the model architecture and reflect a growing trend toward unified, multimodal understanding and layout-aware reasoning to improve document comprehension accuracy and versatility.

7 Challenges and Future Direction

Synthetic Data. Acquiring high-quality, manually curated datasets for new document collections is often quite costly. Leveraging synthetically generated datasets offers a cost-effective alternative for adapting to the target domain (Ding et al., 2025a, 2026). For large-scale instruction-tuning, many frameworks generate instruction-response pairs using benchmarks, templates, or LLMs. However, these synthetic datasets often lack validation, resulting in noise. Since synthetic data may not fully capture real user input, future research should prioritize human-in-the-loop and reinforcement learning approaches to improve authenticity and task relevance.

Long Document Understanding. In practice, VRDs frequently span multiple pages; however, most existing frameworks are tailored for single-page inputs. Multi-page approaches typically rely on retrievers to identify relevant pages, which are then processed by MLLM-based VRDU systems. These methods often fall short of capturing semantic and logical dependencies among document entities, resulting in incomplete contextual understanding. Furthermore, handling long input sequences remains challenging, as existing multi-page benchmarks focus mainly on extractive tasks and rarely support complex multi-hop or multi-modal reasoning.

Multilingual VRDU. Most existing models and benchmarks remain heavily English-centric, limiting their generalization to documents with diverse languages and layouts. This bias is further amplified by large-scale pretraining corpora that predominantly reflect English document structures, leading to performance degradation in low-resource settings. Although few multilingual datasets have been proposed (Xu et al., 2022; Chen et al., 2025), future research should explore more multilingual and culturally diverse benchmarks, language-agnostic representation learning, and hybrid approaches to mitigate linguistic bias to handle real-world document diversity.

Effective RAG Framework. While RAG has become a common paradigm (Jain et al., 2025; Zhang et al., 2026; Faysse et al., 2025), existing approaches often exhibit brittle retrieval due to layout ambiguity and misaligned multimodal embeddings, leading to unreliable evidence selection. Moreover, most RAG pipelines decouple retrieval from reasoning and remain largely text-centric, limiting their ability to capture spatial and visual semantics in complex documents. Future work should explore multimodal RAG frameworks that support iterative reasoning and dynamic evidence refinement, and enable more robust and interpretable VRDU.

Agentic LLM in VRDU. Recent works (Han et al., 2025; Sun et al., 2025) incorporate external tools (e.g., PDF parsers or retrievers) to generate intermediate outputs, enhancing both the accuracy and interpretability of practical VRDU applications. However, future research should explore a wider variety of agent types and architectural innovations to enable automatic handling of diverse formats, cross-domain scenarios, and fine-grained elements such as charts and tables. Additionally, challenges in agentic AI, such as multi-agent coordination and knowledge conflicts, remain significant barriers to broader adoption for VRDU.

Limitations

While this survey offers a comprehensive overview of MLLM-based VRDU research, our analysis is necessarily qualitative. It does not provide exhaustive head-to-head comparisons, as the field's rapid evolution and breadth prioritize trend summarization over detailed benchmarking. Although academic advances are thoroughly reviewed, discussion of real-world deployments and industrial challenges remains limited, in part because many practical applications are proprietary and unpublished. In future work, we aim to provide more quantitative meta-analyses, incorporate insights from industrial adoption, and continuously update the survey to capture the latest developments as the field progresses.

Acknowledgements

This research was supported by the Australian Research Council (ARC) Training Centre for Critical Resources for the Future (CCRF) under grant number IC230100035. This work was also supported by the National Library of Medicine [grant numbers R01LM014344, R01LM014573] and the National Science Foundation (NSF) [grant numbers 2145640, 2139899].

References

Camille Barboule, Benjamin Piwowarski, and Yoan Chabot. 2025. Survey on question answering over visually rich documents: Methods, challenges, and trends. *arXiv preprint arXiv:2501.02235*.

Galal M Binmakhashen and Sabri A Mahmoud. 2019. Document layout analysis: a comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618.

Ketong Chen, Yuhao Chen, and Yang Xue. 2025. Mosaicdoc: A large-scale bilingual benchmark for visually rich document understanding. *arXiv preprint arXiv:2511.09919*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Julien Delaunay, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. 2023. A comprehensive survey of document-level relation extraction (2016-2023). *arXiv preprint arXiv:2309.16396*.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. 2025. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1135–1159, Vienna, Austria. Association for Computational Linguistics.

Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. In *Workshop on Document Intelligence at NeurIPS 2019*.

Yihao Ding, Soyeon Caren Han, Yanbei Jiang, Yan Li, Zechuan Li, and Yifan Peng. 2025a. Syndoc: A hybrid discriminative-generative framework for enhancing synthetic domain-adaptive document key information extraction. *arXiv preprint arXiv:2509.23273*.

Yihao Ding, Soyeon Caren Han, Jean Lee, and Eduard Hovy. 2025b. Deep learning based visually rich document content understanding: A survey. *arXiv preprint arXiv:2408.01287*.

Yihao Ding, Siqu Long, Jiabin Huang, Kaixuan Ren, Xingxiang Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. Form-nlu: Dataset for the form natural language understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2807–2816. ACM.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024a. Mmvqa: A comprehensive dataset for investigating multipage multimodal information retrieval in pdf-based visual question answering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI*, pages 3–9. ijcai.org.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024b. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

Yihao Ding, Qiang Sun, Puzhen Wu, Sirui Li, Siwen Luo, and Wei Liu. 2026. Docs2synth: A synthetic data trained retriever framework for scanned visually rich documents understanding. *arXiv preprint arXiv:2601.12260*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, and 1 others. 2020.

An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *ICLR*.

Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*, 67(12):1–14.

Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. 2023. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*.

Tongkun Guan, Zining Wang, Pei Fu, Zhengtao Guo, Wei Shen, Kai Zhou, Tiezhu Yue, Chen Duan, Hao Sun, Qianyi Jiang, and 1 others. 2025. A token-level text image foundation model for document understanding. *arXiv preprint arXiv:2503.02304*.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *Preprint*, arXiv:2503.13964.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19485–19494. IEEE.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10767–10775.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5817–5834.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091. ACM.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Chelsi Jain, Yiran Wu, Yifan Zeng, Jiale Liu, Shengyu Dai, Zhenwen Shao, Qingyun Wu, and Huazheng Wang. 2025. SimpleDoc: Multi-Modal document understanding with Dual-Cue page retrieval and iterative refinement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28398–28415, Suzhou, China. Association for Computational Linguistics.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4459–4469. Association for Computational Linguistics.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 498–517. Springer.

Sungnyun Kim, Haofu Liao, Srikar Appalaraju, Peng Tang, Zhuowen Tu, Ravi Kumar Satzoda, R Manmatha, Vijay Mahadevan, and Stefano Soatto. 2024. Dockd: Knowledge distillation from llms for open-world document understanding models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3167–3193. Association for Computational Linguistics.

Marcel Lamott, Yves-Noel Weweler, Adrian Ulges, Faisal Shafait, Dirk Krechel, and Darko Obradovic.

2024. Lapdoc: Layout-aware prompting for documents. In *International Conference on Document Analysis and Recognition*, pages 142–159. Springer.

David D Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and James Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*.

Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. 2024. Enhancing visual document understanding with contrastive learning in large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15546–15555.

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. 2024. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*.

Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. 2024a. Hrvda: High-resolution visual document assistant. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15534–15545.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024c. Textmonkey: An ocr-free large multimodal model for understanding document. *Preprint*, arXiv:2403.04473.

Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, and 1 others. 2024. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 15630–15640. IEEE.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, and 1 others. 2023. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*.

Pengyuan Lyu, Yulin Li, Hao Zhou, Weihong Ma, Xingyu Wan, Qunyi Xie, Liang Wu, Chengquan Zhang, Kun Yao, Errui Ding, and 1 others. 2024. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond. *arXiv preprint arXiv:2405.21013*.

Ahmed Masry, Juan A Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-Andre Noel, and 1 others. 2025. Alignvlm: Bridging vision and language latent spaces for multimodal document understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Oshri Naparstek, Roi Pony, Inbar Shapira, Foad Abo Dahood, Ophir Azulai, Yevgeny Yaroker, Nadav Rubinstein, Maksym Lysak, Peter Staar, Ahmed Nassar, Nikolaos Livathinos, Christoph Auer, Elad Amrani, Idan Friedman, Orit Prince, Yevgeny Burshtein, Adi Raz Goldfarb, and Udi Barzelay. 2024. Kvp10k : A comprehensive dataset for key-value pair extraction in business documents. *Preprint*, arXiv:2405.00505.

Feng Ni, Kui Huang, Yao Lu, Wenyu Lv, Guanzhong Wang, Zeyu Chen, and Yi Liu. 2025. Pp-docbee: Improving multimodal document understanding through a bag of tricks. *Preprint*, arXiv:2503.04065.

Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. 2024. Jdocqa: Japanese document question answering dataset for generative language models. *Preprint*, arXiv:2403.19454.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jaeyoo Park, Jin Young Choi, Jeonghyung Park, and Bohyung Han. 2024. Hierarchical visual feature aggregation for ocr-free document understanding. *Advances in Neural Information Processing Systems*, 37:105972–105996.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, and 1 others. 2024. Lmdx: Language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15140–15168.

Minenobu Seki, Masakazu Fujio, Takeshi Nagasaki, Hiroshi Shinjo, and Katsumi Marukawa. 2007. Information management system using structure analysis of paper/electronic documents and its applications. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 689–693. IEEE.

Yufan Shen, Chuwei Luo, Zhaoqing Zhu, Yang Chen, Qi Zheng, Zhi Yu, Jiajun Bu, and Cong Yao. 2025. Proctag: Process tagging for assessing the efficacy of document instruction data. *Preprint*, arXiv:2407.12358.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. 2023. DocILE benchmark for document information localization and extraction.

Li Sun, Liu He, Shuyue Jia, Yangfan He, and Chenyu You. 2025. Docagent: An agentic framework for multi-modal long-context document understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17712–17727.

Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 19071–19079. AAAI Press.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yangfan He, Kuan Lu, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2025. Mtvqa: Benchmarking multilingual text-centric visual question answering. *Preprint*, arXiv:2405.11985.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, and 1 others. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. Docllm: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. 2023. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. 2025a. Mmlongbench: Benchmarking long-context vision-language models effectively and thoroughly. *Preprint*, arXiv:2505.10610.

Zining Wang, Tongkun Guan, Pei Fu, Chen Duan, Qianyi Jiang, Zhentao Guo, Shan Guo, Junfeng Luo, Wei Shen, and Xiaokang Yang. 2025b. Marten: Visual question answering with mask generation for multi-modal document understanding. *Preprint*, arXiv:2503.14140.

Toyohide Watanabe, Qin Luo, and Noboru Sugie. 1995. Layout recognition of multi-kinds of table-form documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4):432–445.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer.

Biao Xiang, Soyeon Caren Han, and Yihao Ding. 2026. Bridge: Benchmark for multi-hop reasoning in long multimodal documents with grounded evidence. *arXiv preprint arXiv:2603.07931*.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. 2025.

Seephys: Does seeing help thinking? – benchmarking vision-based physics reasoning. *Preprint*, arXiv:2505.19099.

Xudong Xie, Hao Yan, Liang Yin, Yang Liu, Jing Ding, Minghui Liao, Yuliang Liu, Wei Chen, and Xiang Bai. 2025. Pdf-wukong: A large multimodal model for efficient long pdf reading with end-to-end sparse sampling. *Preprint*, arXiv:2410.05970.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, and 1 others. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200. ACM.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. Xfund: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the association for computational linguistics: ACL 2022*, pages 3214–3224.

Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. 2017. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324. IEEE Computer Society.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. 2024. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *Preprint*, arXiv:2412.02210.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *Preprint*, arXiv:2307.02499.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, and 1 others. 2023b. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023c. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Wenwen Yu, Zhibo Yang, Yuliang Liu, and Xiang Bai. 2025. Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 837–847.

Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. 2024a. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *Preprint*, arXiv:2404.09204.

Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. 2024b. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *Preprint*, arXiv:2410.05261.

Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2024a. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101*.

Jinxu Zhang, Qiyuan Fan, and Yu Zhang. 2025. Docassistant: Integrating key-region reading and stepwise reasoning for robust document visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3496–3511.

Renshan Zhang, Yibo Lyu, Rui Shao, Gongwei Chen, Weili Guan, and Liqiang Nie. 2024b. Token-level correlation-guided compression for efficient multimodal document understanding. *arXiv*.

Ruiyi Zhang, Yufan Zhou, Jian Chen, Jiuxiang Gu, Changyou Chen, and Tong Sun. 2024c. Llava-read: Enhancing reading ability of multimodal language models. *arXiv preprint arXiv:2407.19185*.

Wenxiao Zhang, Yu Liu, Yihao Ding, Sirui Li, Yanbing Liu, Jin B Hong, Wei Liu, and 1 others. 2026. Stindex: A context-aware multi-dimensional spatiotemporal information extraction system. *arXiv preprint arXiv:2604.08597*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2024d. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *Preprint*, arXiv:2306.17107.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

Yuke Zhu, Yue Zhang, Dongdong Liu, Chi Xie, Zihua Xiong, Bo Zheng, and Sheng Guo. 2025a. Enhancing document understanding with group position embedding: A novel approach to incorporate layout information. In *The Thirteenth International Conference on Learning Representations*.

Zhaoqing Zhu, Chuwei Luo, Zirui Shao, Feiyu Gao, Hangdi Xing, Qi Zheng, and Ji Zhang. 2025b. A simple yet effective layout token in large language models for document understanding. *arXiv preprint arXiv:2503.18434*.

A More Framework Details

A.1 Open-source Frameworks

Table 2 presents official open-source links for VRDU and MLLM frameworks, underscoring the vital role of open access in fostering transparency, reproducibility, and accelerated innovation within the research community.

A.2 Model Training Paradigm Comparison

Table 3 provides a comprehensive comparison of MLLM-based VRDU frameworks across three major training stages: Pretraining (PT), Instruction-tuning (IT), and Supervised Fine-tuning (SFT). OCR-dependent models generally rely on external text extraction and have limited pretraining because they are trained on OCR-processed inputs. In contrast, OCR-free models, which operate directly on document images, demonstrate richer instruction-tuning and fine-tuning strategies, often involving frozen or LoRA-based vision and language encoders. This highlights the diverse training paradigms and modular designs adopted to balance efficiency, adaptability, and performance across frameworks.

A.3 Model Component Details

Table 4 presents a comprehensive comparison of component configurations adopted by recent MLLM-based frameworks for VRDU, spanning both OCR-Dependent and OCR-Free paradigms. For each model, we summarize its LLM backbone (e.g., Vicuna, Qwen, LLaMA, GPT), vision encoder (e.g., CLIP, ViT, Swin), input resolution (including dynamic scaling and cropping), and specialized adaptors or projectors (e.g., LoRA, MLP, QPN) used for multimodal fusion. OCR-Dependent models typically incorporate layout-aware encoders (e.g., LayoutLMv3, DocFormer) and rely on structured textual inputs. In contrast, OCR-Free models process raw document images directly, often requiring higher resolutions and additional modules such as resamplers, visual abstractors, or cropping strategies. The table also lists the maximum supported image resolution, indicating each model's capacity for fine-grained visual understanding. This comparison highlights the increasing diversity in MLLM architectures and the adoption of lightweight tuning techniques for scalable VRDU.

A.4 Document Parsing Tools

Table 5 provides a comparative overview of representative OCR engines, document parsing APIs, and vision–language models for document understanding. The table highlights clear trade-offs across deployment modes, pricing models, and functional capabilities: traditional OCR engines are predominantly open-source and locally deployable but offer limited support for structured document parsing, while commercial document APIs and vision LLMs more frequently provide GPU acceleration and native document-structure extraction at the cost of cloud dependency and usage-based pricing. Recent vision–language models bridge OCR and higher-level reasoning by supporting multimodal inputs (image and PDF) and multilingual processing, yet vary substantially in openness and deployment flexibility. Overall, the comparison illustrates the evolving landscape from text-centric OCR toward multimodal, structure-aware document understanding systems.

B Dataset Overview.

Tables 6 and 7 summarize the datasets used across different training stages. Pretraining typically relies on large-scale, cross-domain document corpora (e.g., IIT-CDIP, RVL-CDIP) to build general multimodal understanding, sometimes extended with domain-specific collections. Instruction-tuning datasets are constructed either from benchmark datasets or via synthetic generation to improve instruction following and domain adaptation. For downstream optimization, supervised fine-tuning commonly leverages QA-style benchmarks (e.g., DocVQA, MPDocVQA) and reformulates key information extraction datasets (e.g., FUNSD, CORD). These tables provide a structured overview of dataset sources and their roles in the training pipeline.

C Quantitive Analysis

C.1 Performance on Single Page Benchmarks

Table 9 highlights clear trends in the performance of general-domain LLMs/MLLM and OCR-dependent and OCR-free document understanding frameworks across several popular benchmarks. Generally, OCR-dependent models achieve consistently strong results on classic form and receipt datasets such as FUNSD, CORD, and SROIE—often exceeding 80% accuracy, with

Framework	Model Name	Official Open Source Link
mPLUG-DocOwl 1.5	DocOwl 1.5	github.com/X-PLUG/mPLUG-DocOwl/tree/main/DocOwl1.5
mPLUG-DocOwl 2	DocOwl 2	github.com/X-PLUG/mPLUG-DocOwl/tree/main/DocOwl2
UReader	UReader	github.com/X-PLUG/mPLUG-DocOwl/tree/main/UReader
KOSMOS-2.5	KOSMOS-2.5 / 2.5-CHAT	aka.ms/kosmos25
LLaVAR	LLaVAR	github.com/SALT-NLP/LLaVAR
Marten	Marten	github.com/PriNing/Marten
LEOPARD	LEOPARD	github.com/Jill0001/Leopard

Table 2: Official open-source links for some VRDU/MLLM frameworks.

top models such as PDF-WuKong, GPE, and DocLayLLM achieving state-of-the-art performance. In contrast, OCR-free frameworks, while demonstrating rapid progress, still lag on these traditional datasets but show remarkable advances on more visually and semantically complex benchmarks such as DocVQA, ChartVQA, and InfoVQA. Notably, the latest OCR-free models, including Texthawk2, Marten, and PP-DocBee, have begun to outperform or match OCR-dependent methods on DocVQA and chart-centric tasks, signaling a narrowing of the gap in real-world document reasoning capabilities. However, coverage remains uneven, with many OCR-free models performing poorly on specific datasets, indicating ongoing challenges with generalizability and benchmark saturation. Overall, while OCR-dependent methods remain dominant for structured text extraction, OCR-free approaches are quickly maturing and expanding the frontier of end-to-end document understanding.

C.2 Performance on Multi-Page Benchmarks

We report the performance of existing multi-page frameworks on two multi-page VRDU benchmarks in Table 10. General-domain models can achieve reasonable performance; however, frameworks equipped with mechanisms explicitly designed for visually rich documents (VRDs) consistently yield substantial improvements. Currently, most high-performing multi-page methods rely on OCR-dependent pipelines and achieve strong results by leveraging external OCR tools. While such designs reduce the burden of directly understanding and compressing visual representations, they also inherit the limitations of OCR-based approaches, including error accumulation as observed in single-page scenarios. For multi-page tasks, this challenge is further amplified, highlighting the need for more effective strategies to manage the large number of visual tokens and to improve text understanding in multi-page, text-dense document inputs.

Model Name	Vision Encoder PT	Vision Encoder IT	Vision Encoder SFT	LLM Backbone PT	LLM Backbone IT	LLM Backbone SFT	Adaptors PT	Adaptors IT	Adaptors SFT
OCR-Dependent									
ICL-D3IE (2023)	–	–	–	–	–	–	–	–	–
DocLLM (2024a)	✓	✓	–	–	–	–	✓	✓	–
LAPDoc (2024)	–	–	–	–	–	–	–	–	–
LMDX (2024)	–	–	–	–	–	–	–	–	–
ProcTag (2025)	–	–	✓	–	–	✓	–	–	✓
DocKD (2024)	–	–	✓	–	–	✓	–	–	–
DoCo (2024)	✗	–	✗	✓	–	✗	✓	–	✓
InstructDoc (2024)	–	✗	✗	–	✗	✗	–	✓	✓
LayoutLLM (2024)	–	✗	✓	–	✗	✗	–	✓	✓
LLaVA-Read (2024c)	✗	✓	–	✗	✗	–	✓	✓	–
LayTextLLM (2024)	✗	–	✓	–	–	–	✓	–	✓
LayTokenLLM (2025b)	✗	–	✗	–	–	–	✓	–	✓
GPE (2025a)	–	–	✓	–	–	–	–	–	–
MDocAgent (2025)	–	–	–	–	–	–	–	–	–
PDF-WuKong (2025)	–	–	✓	–	–	✓	–	–	–
DocLayLLM (2024)	✗	✗	–	✓	✓	–	✓	✓	–
DocAssistant (2025)	-	✗	-	-	✗	-	-	✓	-
AlignVLM (2025)	✓	✓	✗	✓	✓	✓	✓	✓	✓
DocThinker (2025)	-	-	✓	-	-	✓	-	-	✓
OCR-Free									
KOSMOS-2.5 (2023)	–	✓	✓	–	✓	✗	–	✓	✓
mPLUG-DocOwl (2023a)	–	✗	–	–	✗	–	–	✓	–
UReader (2023b)	–	✗	–	–	✗	–	–	✓	–
TGDoc (2023)	–	✗	✓	–	✗	✗	–	✓	✓
UniDoc (2023)	–	✗	✓	–	✗	✗	–	✓	✓
DocPedia (2024)	✗	–	✓	✓	–	✓	✓	–	✓
HRVDA (2024a)	✗	✗	–	✓	✗	–	✓	✓	–
Vary (2024)	✓	–	✓	✓	–	✗	✓	–	✓
mPLUG-DocOwl 1.5 (2024)	–	✗	✓	–	✓	✗	–	✓	✓
HVFA (2024)	–	✗	–	–	✗	–	–	✓	–
mPLUG-DocOwl2 (2025)	–	✗	✓	–	✓	✗	–	✓	✓
Texthawk (2024a)	–	✗	✓	–	✗	✗	–	✓	✓
Texthawk2 (2024b)	–	✗	✓	–	✗	✓	–	✓	✓
TextMonkey (2024c)	–	✓	–	–	✓	–	✓	✓	–
Llavar (2024d)	–	✗	✓	–	✗	✗	–	✓	✓
TokenCorrCompressor (2024b)	–	–	✗	–	–	✗	–	–	✓
DocKylin (2024a)	–	✗	✓	–	✓	✓	–	✓	✓
Marten (2025b)	–	✗	✓	–	✓	✓	–	✓	✓
PP-DocBee (2025)	–	–	✓	–	–	✗	–	–	–
TokenFD (2025)	✓	✗	✓	✓	✗	✓	✓	✓	✓

Table 3: Comparison of MLLM-based VRDU frameworks. PT - Pretraining, IT - Instruction-tuning, SFT - Supervised Fine-tuning.

Model	LLM Backbone	Vision	Adaptor
OCR-Dependent			
ICL-D3IE	GPT-3, ChatGPT	–	–
DocLLM	Falcon-1B / LLaMA2-7B	–	Spatial Attention
LAPDoc	ChatGPT, Solar	–	–
LMDX	PaLM2-S, Gemini Pro	–	–
ProcTag	Qwen-7B / Qwen-VL	Qwen2VL	Projector
DocKD	DocFormerV2	DocFormerV2	–
DoCo	Qwen-VL / mPLUG-Owl	ViT-bigG	VL Adapter
InstructDr	Flan-T5	CLIP	DocFormer
LayoutLLM	Vicuna / LLaMA2	LayoutLMv3	MLP
LLaVA-Read	Vicuna-13B	CLIP-ViT-L	MLP
LayTextLLM	LLaMA2-7B	–	Layout LoRA
LayTokenLLM	Qwen / LLaMA3	–	Layout Tokenizer
GPE	LLaMA2 / Qwen	–	–
MDocAgent	LLaMA3 / Qwen2-VL	ColPali	–
PDF-WuKong	IXC2-VL	IXC2-VL	–
DocLayLLM	LLaMA2 / LLaMA3	LayoutLMv3	Projector + LoRA
DocAssistant	InternVL2	InternVL2	MoM Adapter
AlignVLM	Llama3.1	SigLIP	ALIGN
DocThinker	Qwen2.5-VL	Qwen2.5-VL	–
OCR-Free			
KOSMOS-2.5	Transformer	Pix2Struct	Resampler
DocOwl	mPLUG-Owl	ViT	Abstractor
UReader	mPLUG-Owl	CLIP-ViT	Abstractor
TGDoc	Vicuna-7B	CLIP	MLP
UniDoc	Vicuna	CLIP	MLP
DocPedia	Vicuna-7B	Swin	MLP
HRVDA	LLaMA2	Swin	Detector + LoRA
Vary	OPT + Qwen	CLIP + SAM	MLP
DocOwl1.5	mPLUG-Owl2	ViT	Reducer
HVFA	BLIP2 / mPLUG	ViT	HVFA + LoRA
DocOwl2	mPLUG-Owl2	ViT	Reducer
Texthawk	InternLM	SigLIP	Resampler
Texthawk2	Qwen2	SigLIP	Multi-module
TextMonkey	Qwen-VL	ViT-BigG	Resampler
LLaVAR	Vicuna-13B	CLIP	MLP
TokenCorr	LLaMA2	CLIP	Compressor
DocKylin	Qwen-7B	Swin	MLP + APS
Marten	InternLM2	InternViT	Mask Module
PP-DocBee	Qwen2-VL	ViT	–
TokenFD	Embedding	ViT	Abstractor

Table 4: MLLM-based VRDU frameworks.

Tool Name	Provider	Tool Type	Deployment	Pricing	Input Modalities	Languages	Openness	GPU	Doc Parsing
pdfminer.six	Y. Shinyama et al.	OCR Engine	Local	Free	PDF	Multi	Open-source	No	No
Mistral OCR	Mistral AI	Document API	Cloud	Paid (Usage-based)	Image, PDF	Multi	Closed	Supported	Yes
LightOnOCR	LightOnAI	Vision LLM	Cloud	Paid (Usage-based)	Image, PDF	Multi	Closed	Supported	No
Google Cloud Vision	Google	Document API	Cloud	Paid (Usage-based)	Image, PDF	Multi	Closed	Supported	Yes
Kraken	Inria et al.	OCR Engine	Local	Free	Image, PDF	Multi	Open-source	Supported	No
Qwen3-VL	Aliyun	Vision LLM	Hybrid	Free*	Image, PDF	Pretrained/Dependent	Closed	Supported	No
olmOCR	AI2	OCR Engine	Hybrid	Free	Image, PDF	Multi	Open-source	Supported	Yes
AttentionOCR	Guo & Deng	OCR Engine	Local	Free	Image	Multi	Open-source	Supported	No
Calamari	Univ. Würzburg	OCR Engine	Local	Free	Image	Multi	Open-source	Supported	No
EasyOCR	JaidedAI	OCR Engine	Local	Free	Image	Multi	Open-source	Supported	No
OpenAI Vision	OpenAI	Vision LLM	Cloud	Paid (Usage-based)	Image, PDF	Multi	Closed	Supported	Yes
Tesseract	S. Weil	OCR Engine	Local	Free	Image	Multi	Open-source	No	No
Adobe PDF Extract	Adobe	Document API	Cloud	Paid (Usage-based)	PDF	Multi	Closed	Supported	Yes
PaddleOCR	PaddlePaddle	OCR Engine	Cloud	Free	Image, PDF	Multi	Open-source	Supported	Yes
docTR	Mindee	OCR Engine	Local	Free	Image, PDF	Pretrained/Dependent	Open-source	Supported	No
DeepSeek-OCR	DeepSeek AI	Vision LLM	Hybrid	Paid (Usage-based)	Image, PDF	Multi	Open-source	Supported	No
HunyuanOCR	Tencent	Vision LLM	Local	Free	Image, PDF	Multi	Open-source	Supported	No
Ocular	Berkeley NLP	OCR Engine	Local	Free	Image, PDF	Multi	Open-source	Supported	No
MinerU	OpenDataLab	Document API	Local	Free	PDF	Multi	Open-source	Supported	Yes
SuryaOCR	Datalab	OCR Engine	Local	Free	Image, PDF, Word, PPT	Multi	Open-source	Supported	No
Seed-VL	ByteDance Seed	Vision LLM	Cloud	Paid (Usage-based)	Image, PDF	Multi	Open-source	Supported	Yes

Table 5: Comparison of OCR engines, document parsing APIs, and vision-language models for document understanding.

Study	Dataset	Source	Size	Public Available
Vary	Document Data Engine	ArXiv, CC-MAIN, E-books	2M	✗
	Chart Data Engine	matplotlib, pyecharts, NLP corpora	1.5M	✗
	Detection Data Engine	Objects365, OpenImages	∼3M	✓
LLaVAR	LAION	LAION images filtered for text-rich content, OCR applied	0.4M	✓
DoCo	DoCo-Processed	CC3M (LLaVA) + LAION, processed with PaddleOCR	1.0M	✗
Texthawk2	100M pretraining	Diverse, mainly public datasets	100M	✗
Docpedia	PDF Images	arXiv (public scientific preprints)	325K	✓
	PPT Images	Common Crawl (web-crawled PPTs)	600K	Partly

Table 6: Summary of pretraining datasets created and used in recent MLLM-based VRDU frameworks.

Framework	Category	Source / Description	Size (K)	Open Source
Leopard	Multi-image (text-rich)	69K public multi-page docs/slides; Adapted single-page to multi-image (DocVQA, ArxivQA); Raw slides + GPT-4o QAs; Multi-chart/table (open, synth.); Webpage snapshots (Mind2Web, OmniACT, WebScreenshots, etc.)	739	Partially
	Single-image	Text-rich single images from public datasets; Natural images (e.g., ShareGPT4V, etc.)	186	Partially
LLaVAR	Noisy Instruction-Following	Text-rich images from LAION, selected via classifier + CLIP clustering, instructions via OCR-based prompts	422,000	Yes
	High-Quality Instruction-Following	Subset of LAION text-rich images (4 clusters), multi-turn QAs generated by prompting text-only GPT-4 with OCR+caption info	16,000	Yes

Table 7: Summary of instruction-tuning datasets for Leopard and LLaVAR.

Dataset	Venue	Year	Domain	Docs	Images	Keys / Qs	Multi page	Language	Metrics	Format
Key Information Extraction										
FUNSD	ICDAR-w	2019	Multi-source	–	199	4	✗	English	F1	P, H
SROIE	ICDAR-c	2019	Scanned Receipts	–	973	4	✗	English	F1*	P
CORD	NeurIPS-w	2019	Scanned Receipts	–	1,000	54	✗	English	F1	P
Payment-Invoice	ACL	2020	Invoice Form	–	14,832	7	✗	English	F1	D
Payment-Receipts	ACL	2020	Scanned Receipts	–	478	2	✗	English	F1	P
Kleister-NDA	ICDAR	2021	Private Agreements	540	3,229	4	✓	English	F1	D
Kleister-Charity	ICDAR	2021	AFR	2,778	61,643	8	✓	English	F1	D, P
EPHOIE	AAAI	2021	Exam Paper	–	1,494	10	✗	Chinese	F1	P, H
XFUND	ACL	2022	Synthetic Forms	–	1,393	4	✗	Multilingual	F1	D, P, H
Form-NLU	SIGIR	2023	Financial Form	–	857	12	✗	English	F1	D, P, H
VRDU-Regist. Form	KDD	2023	Registration Form	–	1,915	6	✗	English	F1	D
VRDU-Ad-buy Form	KDD	2023	Political Invoice Form	–	641	9+1(5)	✗	English	F1	D, P
DocILE	ICDAR	2023	Invoice Form	6,680	106,680	55	✓	English	AP, CLEval	D, P
KVP10k	ICDAR	2024	Cross-domain	–	10,707	118,868	✗	English	F1, IOU	D, H
CC-OCR-KIE	ICCV	2025	Cross-domain	–	2,008	34(-)	✗	Multilingual	F1	D, P, H
Visual Question Answering										
DocVQA	WACV	2021	Industrial Reports	–	12,767	50,000	✗	English	ANLS	D, P, H
VisualMRC	AAAI	2021	Website	–	10,197	30,562	✗	English	BLEU, etc	D
TAT-DQA	MM	2022	Financial Reports	2,758	3,067	16,558	✓	English	EM, F1	D
RDVQA	MM	2022	Data Analysis Report	8,362	8,514	41,378	✗	English	ANLS, ACC	D
CS-DVQA	MM	2022	Industry Documents	–	600	1,000	✗	English	ANLS	D, P, H
InfographicVQA	WACV	2022	Infographics	–	5,400	3,000	✗	English	ANLS, F1	D
PDFVQA-Task A	ECML-PKDD	2023	Academic Paper	–	12,337	81,085	✗	English	F1	D
PDFVQA-Task B	ECML-PKDD	2023	Academic Paper	–	12,337	53,872	✗	English	F1	D
PDFVQA-Task C	ECML-PKDD	2023	Academic Paper	1,147	12,337	5,653	✓	English	EM	D
MPDocVQA	PR	2023	Industrial Reports	6,000	48,000	46,000	✓	English	ANLS	D, P, H
DUDE	ICCV	2023	Cross-domain	5,019	28,709	41,541	✓	English	ANLS	D
SlideVQA	AAAI	2023	Slide, decks	–	5,200	14,500	✓	English	EM, F1	D
MMLONGBENCH-DOC	NIPS	2024	Cross-domain	135	6,413	1,082	✓	English	ACC, F1	D
MMVQA	IJCAI	2024	Academic Paper	3,146	30,239	262,928	✓	English	EM, PM, MR	D
JDocQA	LREC-COLING	2024	Cross-Domain	5,504	268,000	11,600	✓	Japanese	F1	D
BoundingDocs	IJDAR	2025	Cross-domain, Mixed	48,151	237,437	249,016	✗	Multilingual	ANLS	D, P, H
LongDocURL	ACL	2025	Cross-domain	396	33,000	2,325	✓	English	F1	D
MMDocIR	EMNLP	2025	Cross-domain	6,878	224,223	73,843	✓	Multilingual	F1	D
MTVQA	EMNLP	2025	Cross-domain	-	8,794	28,607	✗	Multilingual	ANLS	D, P, H
SEEPHYS	NIPS	2025	Physics	-	2,245	2,000	✗	English	Accuracy	D

Table 8: Benchmark datasets for Key Information Extraction and Visual Question Answering in visually rich documents. P - Scanned **P**rinted, H - Scanned **H**andwritten, D - **Di**gtial Born

Model Name	FUNSD	CORD	SROIE	DocVQA	ChartVQA	InfoVQA
General Domain LLM						
Qwen1.5-7B-Chat	52.5	29.7	–	64.3	–	–
Llama3-8B-Instruct	57.5	40.0	–	74.2	–	–
General Domain MLLM						
QwenVL-7B	47.1	30.0	–	65.1	–	–
InterVL2-8B	75.8	79.9	–	91.7	–	–
Claude-3.5 Sonnet	–	–	–	88.5	51.8	59.1
GeminiPro-1.5	–	–	–	91.2	34.7	73.9
GPT4o 20240806	–	–	–	92.8	85.7	66.4
OCR-Dependent						
DocLLM (2024a)	51.8	67.4	91.9	69.5	–	–
LAPDoc (2024)	–	–	–	79.8	–	54.9
DoCo (2024)	–	–	–	64.8	68.9	34.9
InstructDr (2024)	38.1	62.7	–	22.3	–	37.6
LayoutLLM (2024)	78.7	62.2	71.0	74.3	–	–
LLaVA-Read (2024c)	36.9	–	58.3	71.0	74.6	36.4
LayTextLLM (2024)	64.0	96.5	95.8	77.2	–	–
LayTokenLLM(2025b)	71.0	75.4	–	85.1	–	–
GPE (2025a)	82.6	86.9	97.8	78.1	–	–
PDF-WuKong (2025)	85.1	–	–	76.9	80.0	61.3
DocLayLLM (2024)	80.7	79.4	84.4	72.8	–	–
AlignVLM (2025)	–	–	–	81.2	75.0	53.8
DocAssistant (2025)	–	–	–	89.8	81.4	66.7
DocThinker (2025)	-	–	81.4	80.2	–	69.7
OCR-Free						
KOSMOS-2.5 (2023)	–	–	–	81.1	62.3	41.3
mPLUG-DocOwl (2025)	–	–	–	62.2	57.4	38.2
UReader (2023b)	–	–	–	65.4	59.3	42.2
TGDoc (2023)	1.7	–	3.0	9.0	11.7	12.8
UniDoc (2023)	1.2	–	1.4	6.5	10.5	13.8
DocPedia (2024)	40.1	–	57.7	49.3	47.8	15.5
HRVDA (2024a)	–	89.3	89.3	91.0	72.1	43.5
Vary-base (2024)	–	–	–	76.3	66.1	–
mPLUG-DocOwl 1.5 (2024)	–	–	–	81.6	70.5	50.4
HVFA (2024)	–	–	–	72.7	63.3	45.9
mPLUG-DocOwl2 (2025)	–	–	–	80.7	70.0	46.4
Texthawk (2024a)	–	–	–	76.4	66.6	50.6
Texthawk2 (2024b)	–	–	–	89.6	81.4	67.8
TextMonkey (2024c)	65.5	67.5	47.0	73.0	66.9	28.6
Llavar-7B (2024d)	1.7	13.6	2.4	11.6	–	–
TokenCorrCompressor (2024b)	–	–	–	78.3	68.9	50.2
DocKylin (2024a)	25.5	–	49.5	77.3	66.8	46.6
Marten (2025b)	44.4	–	80.4	92.0	81.7	75.2
PP-DocBee (2025)	–	–	–	90.6	74.6	66.2
TokenFD (2025)	42.2	–	81.9	94.2	86.6	76.5

Table 9: Performance comparison between OCR-dependent and OCR-free document understanding frameworks across benchmark datasets.

Model	Type	Venue	Year	MPDocVQA	DUDE
Longformer	General VLPM	Preprint	2020	55.1	20.3
BigBird	General VLPM	NeurIPS	2020	58.5	26.3
GPT-4v	General MLLM	–	2023	–	53.9
Idefics3-8B	General MLLM	Preprint	2024	67.2	38.7
LLaVA-next-interleave-7B	General MLLM	Preprint	2024	44.9	28.0
Hi-VT5	OCR-dependent VLPM	PR	2023	61.8	35.7
GRAM	OCR-dependent VLPM	CVPR	2024	**83.0**	53.4
InstructDoc	OCR-Dependent VLPM	AAAI	2024	–	46.8
mPLUG-DocOwl2	OCR-free VLPM	Preprint	2024	69.4	46.7
PDF-WuKong	OCR-Dependent VLPM	Preprint	2024	76.9	56.1
LayTokenLLM	OCR-Dependent VLPM	CVPR	2025	74.3	52.0
DocThinker (2025)	OCR-Dependent VLPM	ICCV	2025	–	**56.8**

Table 10: Performance comparison of state-of-the-art models on MPDocVQA and DUDE benchmarks. Best scores are highlighted in red.