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

Anonymous ACL submission

Abstract

Visually-Rich Document Understanding (VRDU) has emerged as a critical field, driven by the need to automatically process documents containing complex visual, textual, and layout information. Recently, Multimodal Large Language Models (MLLMs) have shown remarkable potential in this domain, leveraging both Optical Character Recognition (OCR)-dependent and OCR-free frameworks to extract and interpret information in document images. This survey reviews recent advancements in MLLM-based VRDU, highlighting three core components: (1) methods for encoding and fusing textual, visual, and layout features; (2) training paradigms, including pretraining strategies, instruction-response tuning, and the trainability of different model modules; and (3) datasets utilized for pretraining, instruction-tuning, and supervised fine-tuning. Finally, we discuss the challenges and opportunities in this evolving field and propose future directions to advance the efficiency, generalizability, and robustness of VRDU systems.

1 Introduction

Visually-Rich Document Understanding (VRDU) sits at the intersection of Natural Language Processing, Computer Vision, and Document Analysis, 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 and generalizable VRDU frameworks has garnered significant attention for automating information extraction, improving accessibility, and enhancing decision-making across diverse domains like 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 are suboptimal when processing unseen documents from different domains or with varied layouts. In contrast, conventional deep learning approaches employed CNNs (Katti et al., 2018; Yang et al., 2017) and RNNs (Denk and Reisswig, 2019) to separately encode visual and textual modalities, providing more flexible representations. However, these methods typically lacked the ability to effectively integrate the diverse modalities in documents, limiting their capacity to capture the rich semantic structure inherent in visually complex documents. With the success of pretraining techniques in a variety of NLP tasks, many VRDU models (Huang et al., 2022; Hong et al., 2022; Lyu et al., 2024) have since been pretrained on large-scale scanned or PDF documents, allowing for more comprehensive multimodal feature fusion. However, their effectiveness is constrained by the scope and diversity of their pretraining data, often requiring substantial fine-tuning for cross-domain generalizability.

Recently, Multimodal Large Language Models (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-rich images with diverse visual appearances and complex spatial layouts. By combining the superior textual understanding of LLMs (Bai et al., 2023) with visual encoders (Dosovitskiy et al., 2020) that capture image content and layout information, MLLM-based VRDU frameworks have shown prominent performances on diverse document structure recognition, information extraction and the generalizability across domains without extensive task-specific fine-tuning, which pushes the boundaries of VRDU.

This paper aims to provide a comprehensive survey on the recent development of MLLM-based VRDU frameworks. Previous surveys either focused on a broad analysis of the diverse capabilities of MLLMs (Caffagni et al., 2024) or the tech-

Model	Venue	Tasks	Mod.	LLM	Vision Enc.	PT	IT	FT	Pages	Prompt In.
OCR-Dependent										
ICL-D3IE (2023)	ICCV	KIE	T, L	GPT-3	N/A	✗	✗	✗	SP	ICL+Layout
DocLLM (2024a)	ACL	KIE, QA, DC	T, L	Custom	N/A	✓	✓	✗	SP	T+B+Q
LAPDoc (2024)	ICDAR	KIE, QA	T, L	Multiple	N/A	✗	✗	✗	SP	Rule
LMDX (2024)	ACL	KIE	T, L	Gemini-pro	N/A	✗	✗	✗	SP	ICL+Layout
ProcTag (2025)	AAAI	QA	T, L	GPT-3.5	N/A	✗	✗	✓	SP	Rule+CoT
DocKD (2024)	EMNLP	KIE, QA, DC	T, L	N/A	N/A	✗	✗	✓	SP	Gen by VL
DoCo (2024)	CVPR	KIE, QA, DC	T, L	Multiple	LayoutLMv3	✓	✗	✓	SP	I+Q
InstructDr (2024)	AAAI	KIE, QA	T, V, L	FlanT5	LayoutLMv3	✗	✓	✓	MP	I+Q
LayoutLLM (2024)	CVPR	KIE, QA	T, V, L	Vicuna-7B-v1.5	OpenCLIP+CLIP	✗	✓	✓	SP	I+Q+CoT
LLaVA-Read (2024c)	preprint	KIE, QA	T, V, L	Vicuna-1.5 13B	N/A	✓	✓	✗	SP	I+Q
LayTextLLM (2024)	preprint	QA, KIE	T, L	Llama2-7B-base	N/A	✓	✗	✓	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	N/A	✓	✗	✗	MP	I+Q+L
GPE (2025a)	ICLR	KIE, QA	T, L	Multi	N/A	✗	✗	✓	SP	T+B+Q
MDocAgent (2025)	preprint	QA	T, V	Multiple	IXC2-VL-4KHD	✗	✗	✗	MP	I+Q
PDF-WuKong (2025)	preprint	QA	T, V	BGE-M3	LayoutLMv3	✗	✗	✓	MP	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	Multi	CLIP, ViTDet	✓	✗	✓	SP	I+Q
mPLUG-DocOwl1.5 (2024a)	EMNLP	KIE, QA	V	mPLUG-Owl2	mPLUG-Owl2 VE	✗	✓	✓	SP	I+Q
HVFA (2024)	NIPS	QA, Cap.	V	Multi (BLIP-2, etc.)	ViT/L-14	✗	✓	✗	SP	I+Q
mPLUG-DocOwl2 (2024b)	preprint	KIE, QA	V	mPLUG-Owl2	ViT	✓	✗	✓	MP	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 (2025)	CVPR	QA	V	InterLM2	InternViT-300M	✗	✓	✓	SP	I+Q
PP-DocBee (2025)	preprint	QA	V	Qwen2-VL-2B	ViT	✗	✗	✓	SP	I+Q

Table 1: Comparison of existing MLLM-based VRDU frameworks. 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.

niques applied to a specific task of document understanding, such as document layout analysis (Binmakhashen and Mahmoud, 2019), question answering (Barboule et al., 2025), and relation extraction (Delaunay et al., 2023). 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 paradigms (Sec 2), **Input Representation** (Sec 3), **Training Strategies** (Sec 4), and **Inference Methods** (Sec 5). We also include a comprehensive list of pretraining and instruction-tuning datasets used by MLLM-based VRDU models (Sec 6) and a detailed discussion of the deficiencies of existing methods and the critical analysis of the trend for future development (Sec 7).

2 Model Architecture

Depending on how multimodal feature is acquired, model architectures are typically classified as either OCR-dependent or OCR-free frameworks [1].

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,

[1] See Appendix A to quantitive analysis and more details.

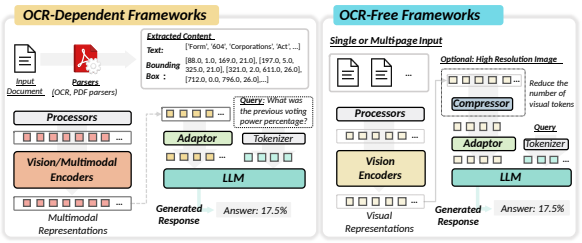

Figure 1: General OCR-dependent and OCR-free framework architectures.

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, mPLUG-DocOwl, and UReader (Kim et al., 2022; Ye et al., 2023a,b). Accurate comprehension of fine-grained textual content in these OCR-free settings requires high-resolution images, which lead to lengthy visual sequences and necessitate visual compression modules (Liu et al., 2024a; Hu et al., 2024a). Moreover, effective text recognition within these models often relies on large-scale pretraining to integrate textual and layout features via tasks such as text recognition (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 for large-scale deployment.

3 Multimodal Representation

This section reviews current trends in how MLLM-based methods encode and fuse textual, visual, and layout features for VRDU (See Figure 2).

3.1 Text Modality

OCR-dependent methods use external tools to extract text, which is then encoded by LLMs or auxiliary encoders. In contrast, OCR-free models take only the document image as input and typically treat text as a training target for integrated learning.

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). 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 extra encoders, but performance is still limited by OCR and LLM errors and weak multimodal integration.

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 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 from document images into the LLM input space via cross- and self-attention. In sum, external encoders improve representations but require extra pretraining and tuning to align with LLMs.

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; Liu et al., 2024a) to integrate text information. To better comprehend document hierarchical structure, Hu et al. (2024a,b) propose a multi-grained text localization task, ranging from the word to the block level. While these methods deliver robust results using only visual

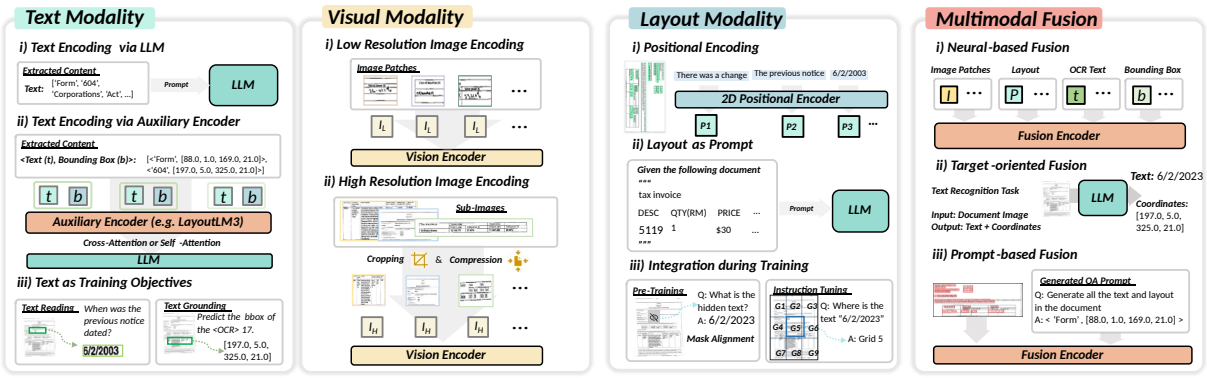

Figure 2: Multimodal feature representation and fusion mechanisms.

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, allowing **lower-resolution** images. In contrast, OCR-free frameworks require direct text recognition, demanding fine-grained perception and **high-resolution** inputs. See the Appendix A.5 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 for multimodal understanding, achieving effective alignment often requires additional pretraining and instruction tuning. Moreover, the lack of fine-grained visual detail typically requires additional OCR tools to extract textual content for accurate VRD understanding.

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. (2024a) splits high-resolution images into predefined sub-images. To handle images with various shapes, UReader (Ye et al., 2023b) introduces a *Shape-Adaptive Cropping Module* to adaptively divide 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 features remains challenging, requiring a balance between resource usage and preserving semantic and layout continuity.

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). For layout-aware compression in visually-rich domains, Hu et al. (2024a) introduces a convolutional module that preserves layout by compressing horizontal features and reducing token count. It further enhances this with layout-aware cross-attention to handle multi-page input. Liu et al. (2024a) incorporates a *Content Detector* that filters non-informative tokens by segmenting text-rich regions, e.g., PSENet (Wang et al., 2019). Zhang et al. (2024a) proposed to eliminate low-information areas and cluster and aggregate 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, combining the text embeddings with 2D positional encodings 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, introducing 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 integrate layout implicitly and depend on large-scale pretraining, reducing effectiveness in tasks needing 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) introduce an in-context learning based approach to incorporate layout-aware demonstrations with 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 textual information 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., 2025; 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 approaches typically require large-scale datasets for pretraining or instruction-tuning, entailing substantial computational resources.

3.4 Multimodal Fusion

After acquiring multimodal features, fusion techniques—categorized as direct, neural-based, prompt-based, and task-oriented—are applied.

Neural-based Fusion. The simplest multimodal feature encoding uses external document encoders such as LayoutLMv3 (Xu et al., 2021), which fuses multimodal features using self or cross-attention with 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 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. 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, especially in noisy OCR scenarios.

Target-oriented Fusion. Target-oriented strategies establish multimodal connections through supervised objectives that span from input to output (Hu et al., 2024a). For example, in text recognition tasks, the model directly learns to map visual features to both textual content and spatial coordinates, thereby aligning feature fusion with the task objective. These approaches enhance end-to-end multimodal integration but introduce higher demands in data preparation, annotation quality, and training complexity in real-world scenarios (Sec 4).

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 step-by-step layout-aware reasoning in document understanding. However, these methods often depend on pre-defined reasoning strategies, intermediate step evaluations, and well-trained preceding frameworks, limiting the generalizability.

4 Training Strategies

To facilitate multimodal understanding, instruction following, and domain adaptation, various training tasks have been developed.

4.1 Pretraining

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., 2024b). OCR-dependent frameworks typically utilize pretrained VRD 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 conduct block-wise text infilling for enhancing enhance text-layout correlation modeling. OCR-free frameworks (Zhang et al., 2024c; Hu et al., 2024a) focus on pretraining tasks like text recognition, detection, and captioning to integrate text and layout information. Hu et al. (2024b) 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 enhance controllability and 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 into instruction-based formats, where images are paired with task descriptions. Beyond improving multimodal fusion, goal-oriented tasks, including VRD question answering (Ding et al., 2024c), 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.2 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, enhancing framework 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 using off-the-shelf OCR tools and LLMs, may yield low-quality QA pairs, particularly impacting zero-shot performance in low-resource domains, such as scanned documents.

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 better interpret 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., 2024a; 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).

Vision/Multimodal Encoders. They are employed to encode multimodal features, which are subsequently aligned with LLM text representations by projectors or adaptors. 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 diverse resolution inputs are often frozen to enhance visual representation from hierarchical inputs. In supervised fine-tuning, there is no standard practice regarding encoder trainability; it is mainly determined empirically.

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., 2024a) 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., 2024b), have also been introduced to reduce the dimensionality of visual features. These adaptors are usually trained during instruction tuning or supervised fine-tuning phases.

5 Inference Prompt Setting

MLLM-based frameworks adopt diverse prompt formats based on their architecture during inference. For OCR-free frameworks in Table 1, the prompt typically includes a document image, occasionally multiple pages (Hu et al., 2024b; Wang et al., 2025), 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 with the query, whereas vision-dependent models also incorporate the document image into vision (Xie et al., 2025) or multimodal encoders (Liao et al., 2024), aligning visual and textual features for 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.

6 MLLM-based VRDU Datasets

Different training stages often utilize distinct datasets to achieve specific objectives. We summarize the datasets employed at each stage, along with a trend analysis. More detailed dataset information can be found in the Appendix B.

Pretraining Datasets. The aim of pretraining is to enable frameworks to enhance multimodal understanding and improve generalizability. Similar to pretrained VRD models (Hong et al., 2022), these frameworks are typically trained on large-scale, cross-domain document collections, such as IIT-CDIP (Lewis et al., 2006), to acquire structural and semantic knowledge across diverse VRD domains. Recent MLLM-based VRDU frameworks (Zhang et al., 2024d; Wang et al., 2023) have introduced new large-scale cross-domain datasets for pretraining. While such datasets improve generalizability, models often struggle with domain-specific tasks. To address this, some frameworks incorporate domain-specific datasets, such as slide decks (Feng et al., 2024), academic papers (Wang et al., 2024a), or downstream task-oriented document collections (Yu et al., 2024b).

Instruction-tuning Datasets. 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 a critical to domain adaptation and accurate instruction interpretation.

Supervised Fine-tuning Datasets. To improve performance on downstream tasks, some frameworks apply supervised fine-tuning on question answering datasets such as DocVQA (Mathew et al., 2021) and MPDocVQA (Tito et al., 2023). Additionally, several key information extraction benchmarks—such as FUNSD (Jaume et al., 2019), FormNLU (Ding et al., 2023), and CORD (Park et al., 2019), have been reformulated into QA-style formats to enable evaluation with generative frameworks. Detailed descriptions of these benchmark datasets are found in Appendix B.3.

7 Discussion and Future Direction

In this survey, we provided a comprehensive review of recent MLLM-based frameworks for VRDU, systematically examining model architectures, multimodal representations, training and inference strategies, and datasets. We identified a clear evolution in design paradigms—from OCR-dependent pipelines to vision-centric, OCR-free approaches, and from isolated modality modelling to more integrated, instruction-driven frameworks. Despite notable progress in VRD-centric tasks, several critical challenges remain underexplored, particularly in real-world applications. These include:

Synthetic Data. In real-world, acquiring high-quality, manually curated datasets for new document collections is often costly. Leveraging synthetically generated datasets, as proposed by Ding et al. (2024a), offers a cost-effective alternative to adapt target domain. 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 low-quality or inaccurate pairs. Verification, particularly with LLMs as evaluators, is crucial to ensure data reliability. 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.

Agent and Retrieval-Augmented Generation. In real-world applications, document-centric tasks are knowledge-intensive and demand a high degree of trustworthiness to ensure reliable responses. Integrating external tools (e.g., PDF parsers, or retrievers) to produce intermediate outputs can enhance response accuracy. Nonetheless, future research should investigate a broader range of agent types and architectural innovations to address diverse formats, cross-domain scenarios, and fine-grained elements like charts and tables.

Long Document Understanding. In real-world scenarios, VRDs frequently span multiple pages; however, most existing frameworks are tailored for single-page inputs (Table 1). 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 in capturing semantic and logical dependencies across document entities, leading to an incomplete contextual understanding. Furthermore, managing long input sequences remains a core challenge for effective retrieval and compression of multimodal content. Existing multi-page benchmarks predominantly target extractive tasks with limited page-spanning complexity, while multi-hop and multimodal reasoning, critical for comprehensive understanding, remain underexplored.

Scaling Law and Domain Adaptation. Most frameworks utilize large-scale datasets for pre-training or instruction-tuning, acquiring general knowledge that boosts performance in zero-shot and fine-tuned settings. Nonetheless, a significant performance gap remains compared to fine-tuned BERT-style VRDU models on domain-specific tasks. While scaling laws suggest performance improves with larger models and data, gains diminish under domain shifts or distributional discrepancies. Despite strong generalizability, large models often falter in domain adaptation due to reliance on heterogeneous corpora that lack fine-grained semantics and layout cues. Their substantial computational and data demands further hinder adaptation in low-resource or specialized domains with limited annotated data. Incorporating diverse agents and synthetic datasets may offer a promising direction to enhance domain adaptation.

Limitations

While this survey offers a comprehensive overview of MLLM-based VRDU research, our analysis is necessarily qualitative and does not provide exhaustive head-to-head comparisons, as the rapid evolution and breadth of the field prioritize trend summarization over detailed benchmarking. Although academic advances are thoroughly reviewed, discussion of real-world deployments and industrial challenges remains limited, in part due to the proprietary and unpublished nature of many practical applications. 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.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

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.

Wenhu Chen, Han Zhu, Wenhao Wang, Kai-Wei Chang, William Yang Zhang, and William Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*.

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*.

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, Zechuan Li, and Hyunsuk Chung. 2024a. David: Domain adaptive visually-rich document understanding with synthetic insights. *arXiv preprint arXiv:2410.01609*.

Yihao Ding, Jean Lee, and Soyeon Caren Han. 2024b. 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. 2024c. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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*.

Pranay Gupta, Minesh Mathew, C.V. Jawahar, and Marcus Liwicki. 2022. Infovqa: Visual question answering on infographics with a multi-modal entity graph. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

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.

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. 2024a. 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. 2024b. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding.

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.

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.

Mengzhao Jia, Wenhao Yu, Kaixin Ma, Tianqing Fang, Zhihan Zhang, Siru Ouyang, Hongming Zhang, Meng Jiang, and Dong Yu. 2024. Leopard: A vision language model for text-rich multi-image tasks. *arXiv preprint arXiv:2410.01744*.

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.

Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, et al. 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, et al. 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, et al. 2024. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond. *arXiv preprint arXiv:2405.21013*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209. IEEE.

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.

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, et al. 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*.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, et al. 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.

Maxim Sidorov, Amanpreet Singh, Yu Li, Jianfeng Liao, Ming Liao, Yaxing Wang, Lichao Wang, Shouling Gong, Chen Change Loy, and Xiang Bai. 2020. Textcaps: A dataset for image captioning with reading. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Amanpreet Singh, Vedanuj Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, Devi Parikh, and Aniruddha Krishnamurthy. 2019. Textvqa: Visual question answering with reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

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*.

Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. 2019. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9336–9345.

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*.

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

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.

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.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 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.

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.

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.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 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, et al. 2023c. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

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.

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.

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*.

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*.

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.

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 Details about MLLM-based VRDU Frameworks

A.1 Performance Comparison

Table 4 highlights clear trends in the performance of OCR-dependent and OCR-free document understanding frameworks across popular benchmarks. Generally, OCR-dependent models achieve consistently strong results on classic form and receipt datasets such as FUNSD, CORD, and SROIE—often surpassing 80% accuracy, with top models like PDF-WuKong, GPE, and DocLayLLM achieving state-of-the-art scores. In contrast, OCR-free frameworks, while demonstrating rapid progress, still lag behind 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 started to outperform or match OCR-dependent methods on DocVQA and chart-centric tasks, signaling a closing gap in real-world document reasoning capabilities. However, coverage remains uneven, with many OCR-free models lacking results on certain datasets, indicating ongoing challenges in 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.

A.2 Framework Key Contributions

Table 5 presents a comparative summary of key techniques underlying recent MLLM-based frameworks for VRDU. Each model employs distinct innovations to enhance multimodal representation and reasoning, spanning novel positional encoding schemes, advanced visual token compression, multi-agent architectures, and instruction-tuning strategies tailored to document tasks such as KIE, QA, and visual grounding. The table highlights how frameworks leverage document structure, spatial layout, and semantic cues through both OCR-dependent and OCR-free paradigms. By distilling the core technical contributions of each approach, this summary provides a comprehensive reference for researchers and practitioners seeking to understand the landscape and evolution of state-of-the-art MLLM-based document understanding solutions.

A.3 Open-source Frameworks

Table 6 presents official open-source links for prominent VRDU and MLLM frameworks, underscoring the vital role of open access in fostering transparency, reproducibility, and accelerated innovation within the research community.

A.4 Model Training Paradigm Comparison

Table 2 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 exhibit limited pretraining due to their reliance 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.5 More Detailed Model Information

Table 3 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 tend to incorporate layout-aware encoders (e.g., LayoutLMv3, DocFormer) and rely on structured textual inputs, while 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.

Frameworks	Model Name	Pretraining			Instruction-tuning			Supervised Fine-tuning		
		PT	IT	SFT	PT	IT	SFT	PT	IT	SFT
OCR-Dependent	ICL-D3IE (2023)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	DocLLM (2024a)	trainable	trainable	N/A	N/A	N/A	N/A	trainable	trainable	N/A
	LAPDoc (2024)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	LMDX (2024)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	ProcTag (2025)	N/A	N/A	trainable	N/A	N/A	trainable	N/A	N/A	trainable
	DocKD (2024)	N/A	N/A	trainable	N/A	N/A	trainable	N/A	N/A	N/A
	DoCo (2024)	frozen	N/A	frozen	trainable	N/A	frozen	trainable	N/A trainable	
	InstructDr (2024)	N/A	frozen	frozen	N/A	frozen	frozen	N/A	trainable	trainable
	LayoutLLM (2024)	N/A	frozen	trainable	N/A	frozen	frozen	N/A	trainable	trainable
	LLaVA-Read (2024c)	frozen	trainable	N/A	frozen	frozen	N/A	trainable	trainable	N/A
	LayTextLLM (2024)	frozen	N/A	trainable	N/A	N/A	N/A	trainable	N/A	trainable
	LayTokenLLM (2025b)	frozen	N/A	frozen	N/A	N/A	N/A	trainable	N/A	trainable
	GPE (2025a)	N/A	N/A	trainable	N/A	N/A	N/A	N/A	N/A	N/A
	MDocAgent (2025)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	PDF-WuKong (2025)	N/A	N/A	trainable	N/A	N/A	trainable	N/A	N/A	N/A
	DocLayLLM (2024)	frozen	frozen	N/A	trainable	trainable	N/A	trainable	trainable	N/A
OCR-Free	KOSMOS-2.5 (2023)	N/A	trainable	trainable	N/A	trainable	frozen	N/A	trainable	trainable
	mPLUG-DocOwl (2023a)	N/A	frozen	N/A	N/A	frozen	N/A	N/A	**trainable***	N/A
	UReader (2023b)	N/A	frozen	N/A	N/A	frozen	N/A	N/A	trainable	N/A
	TGDoc (2023)	N/A	frozen	trainable	N/A	frozen	frozen	N/A	trainable	trainable
	UniDoc (2023)	N/A	frozen	trainable	N/A	frozen	frozen	N/A	trainable	trainable
	DocPedia (2024)	frozen	N/A	trainable	trainable	N/A	trainable	trainable	N/A	trainable
	HRVDA (2024a)	frozen	frozen	N/A	trainable	frozen	N/A	trainable	trainable	N/A
	Vary (2024)	trainable	N/A	trainable	**trainable***	N/A	frozen	trainable	N/A	trainable
	mPLUG-DocOwl 1.5 (2024a)	N/A	frozen	trainable	N/A	trainable	frozen	N/A	trainable	trainable
	HVFA (2024)	N/A	frozen	N/A	N/A	frozen	N/A	N/A	**trainable***	N/A
	mPLUG-DocOwl2 (2024b)	N/A	frozen	trainable	N/A	trainable	frozen	N/A	trainable	trainable
	Texthawk (2024a)	N/A	frozen	trainable	N/A	frozen	frozen	N/A	trainable	trainable
	Texthawk2 (2024b)	N/A	frozen	trainable	N/A	frozen	trainable	N/A	trainable	trainable
	TextMonkey (2024c)	N/A	trainable	N/A	N/A	trainable	N/A	trainable	trainable	N/A
	Llavar (2024d)	N/A	frozen	trainable	N/A	frozen	frozen	N/A	trainable	trainable
	TokenCorrCompressor (2024b)	N/A	N/A	frozen	N/A	N/A	frozen	N/A	N/A	trainable
	DocKylin (2024a)	N/A	frozen	trainable	N/A	trainable	trainable	N/A	trainable	trainable
	Marten (2025)	N/A	frozen	trainable	N/A	trainable	trainable	N/A	trainable	trainable
	PP-DocBee (2025)	N/A	N/A	trainable	N/A	N/A	frozen	N/A	N/A	N/A

Table 2: Comparison of MLLM-based VRDU frameworks: Pretraining, Instruction-tuning, and Supervised Fine-tuning stages. N/A represents no relevant training stage.

B Dataset Details

B.1 Pretraining Dataset

Similar to pretrained VRDU frameworks, MLLM-based approaches also conduct continued pretraining on large-scale document collections to improve multimodal understanding. Commonly used datasets include IIT-CDIP (Lewis et al., 2006), which contains over 6 million scanned documents spanning diverse domains but lacks layout annotations, often supplemented with OCR-derived bounding boxes. RVL-CDIP (Harley et al., 2015) is a smaller, curated subset of 400,000 documents across 16 categories, widely used for classification and low-resource pretraining. Except for those two general domain document collections, some works use self-collected document collections for enhancing specific domains or enriching the document collection types, as summarised by Table 7.

B.2 Instruction Tuning Datasets

As shown by Table 8, some frameworks increasingly generate synthetic instruction-tuning datasets tailored to their architectures, prioritizing alignment over generalizability achieved through benchmark-based tuning.

B.3 Benchmark Datasets

There are a range of benchmark datasets adopted by pretrained VRDU frameworksor MLLM-based frameworks. Table 9 lists the benchmark datasets used for VRD-related Key Information Extraction (KIE) and Visual Question Answering (VQA). Many frameworks are evaluated on other domain-specific datasets as well, including those for chart understanding and webpage analysis. For instance, InfoVQA (Gupta et al., 2022) focuses on visual question answering for information-centric records. Benchmarks like WTQ (Pasupat and Liang, 2015) and TabFact (Chen et al., 2020) assess a model's ability to reason over tabular data, and ChartQA evaluates chart comprehension skills. Additionally, TextVQA (Singh et al., 2019) and TextCaps (Sidorov et al., 2020) target text recognition and semantic reasoning in natural images.

Model Name	LLM Backbone	Vision Backbone	Resolution	Adaptors and Projectors
OCR-Dependent				
ICL-D3IE (2023)	GPT-3 / ChatGPT	N/A	N/A	N/A
DocLLM (2024a)	Falcon-1B/LLaMA2-7B	N/A	N/A	Disentangled Spatial Attention
LAPDoc (2024)	ChatGPT / Solar	N/A	N/A	N/A
LMDX (2024)	PaLM 2-S / Gemini Pro	N/A	N/A	N/A
ProcTag (2025)	Qwen-7B/Qwen-VL-7B	qwen2vl vision encoder	Dynamic (224×224 to 448×448)	qwen2vl projector
DocKD (2024)	DocFormerv2 language decoder	DocFormerv2 vision encoder	Derived from CNN backbone	N/A
DoCo (2024)	Qwen-VL-Chat/mPLUG-Owl	ViT-bigG	224×224	Position-Aware Vision-Language Adapter / Visual Abstractor
InstructDr (2024)	Flan-T5	CLIP	224×224	Document-former
LayoutLLM (2024)	Vicuna-7B-v1.5 / LLaMA2-7B-chat	LayoutLMv3	224×224	MLP
LLaVA-Read (2024c)	Vicuna-1.5 13B	CLIP-ViT-L/14-336 + ConvNext-L/32-320	336×336	MLP
LayTextLLM (2024)	Llama2-7B-base	N/A	320×320	Spatial Layout Projector + Layout Partial LoRA
LayTokenLLM (2025b)	Qwen1.5-7B / LLaMA3-8B	N/A	N/A	Layout Tokenizer + LORA
GPE (2025a)	LLaMA2-7B / Qwen2-7B / ChatGLM-6B	N/A	N/A	N/A
MDocAgent (2025)	LLaMA-3.1-8B (Text), Qwen2-VL-7B (Others)	ColPali	448×448	N/A
PDF-WuKong (2025)	IXC2-VL-4KHD	IXC2-VL-4KHD	Dynamic (336×336 to 3840×1600)	N/A
DocLayLLM (2024)	LLaMA2-7B / LLaMA3-8B	LayoutLMv3 ve	224×224	Layout Embedder + Projector + LORA
OCR-Free				
KOSMOS-2.5 (2023)	Transformer decoder	Pix2Struct-Large ViT-based	1024×1024	Resampler
mPLUG-DocOwl (2023a)	mPLUG-Owl	ViT	224×224	Visual Abstractor + Lora
UReader (2023b)	mPLUG-Owl	CLIP-like ViT	224×224 (×20 crops)	Visual Abstractor + Lora
TGDoc (2023)	Vicuna-7B	CLIP-ViT-L/14	224×224 and 336×336	MLP
UniDoc (2023)	Vicuna	CLIP-ViT-L/14	224×224 and 336×336	MLP
DocPedia (2024)	Vicuna-7B	Swin Transformer	2560×2560	MLP
HRVDA (2024a)	LLaMA-2-7B	Swin Transformer	1536×1536	Content Detector + MLP Projector + LoRA
Vary (2024)	OPT125M + Qwen-7B / Vicuna-7B	CLIP + SAM	1024×1024	MLP
mPLUG-DocOwl 1.5 (2024a)	mPLUG-Owl2	ViT/L-14	448×448 (×9 crops)	H-Reducer
HVFA (2024)	BLIP-2-OPT-2.7B / mPLUG-Owl-7B	ViT	224×224 × crops	HVFA + Lora + Resampler
mPLUG-DocOwl2 (2024b)	mPLUG-Owl2	ViT	504×504 (×12 crops)	H-Reducer
Texthawk (2024a)	InternLM-XComposer 7B	SigLIP-SO (ViT)	224×224 × crops	Resampler + LoRA + QPN + MLCA
Texthawk2 (2024b)	Qwen2-7B-Instruct	SigLIP-SO (ViT)	224×224 × crops (up to 72 crops)	Resampler + QPN + MLCA + Detection Head + LoRA
TextMonkey (2024c)	Qwen-VL-Chat / mPLUG-Owl	ViT-BigG	448×448 × crops	Image + Token Resampler
Llavar (2024d)	Vicuna-13B	CLIP-ViT-L/14	224×224 and 336×336	MLP
TokenCorrCompressor (2024b)	LLaMA2-7B	CLIP-ViT-L/14	224×224 and 336×336	Token Correlation Compressor + LORA
DocKylin (2024a)	Qwen-7B-Chat	Swin (Donut-Swin, 0.07B)	1728×1728	MLP + APS + DTS
Marten (2025)	InternLM2-7B	InternViT-300M	448×448 (×6 crops)	MLP + Mask Generator Module
PP-DocBee (2025)	Qwen2-VL-2B	ViT	1680×1204	N/A

Table 3: Comparison of MLLM-based VRDU frameworks: Backbone and Adapter configurations. "N/A" denotes the component is not applicable or not disclosed.

Model Name	Dataset					
	FUNSD	CORD	SROIE	DocVQA	ChartVQA	InfoVQA
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	-	-
OCR-Free						
KOSMOS-2.5 (2023)	-	-	-	81.1	62.3	41.3
mPLUG-DocOwl (2024b)	-	-	-	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 (2024a)	-	-	-	81.6	70.5	50.4
HVFA (2024)	-	-	-	72.7	63.3	45.9
mPLUG-DocOwl2 (2024b)	-	-	-	80.7	70	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 (2025)	44.4	-	80.4	92.0	81.7	75.2
PP-DocBee (2025)	-	-	-	90.6	74.6	66.2

Table 4: Performance comparison between OCR-dependent and OCR-free document understanding frameworks across benchmark datasets. Dash ('-') indicates that results are not reported.

Model	Key Contribution Summary
LayTokenLLM (2025b)	Introduces layout tokens with a novel positional encoding, enabling spatial awareness without extra IDs and proposing a unique prediction pretraining objective.
DocKD (2024)	Generates synthetic training data with LLMs; encodes layout using markdown-style, spatially linearized OCR text.
DocKylin (2024a)	Proposes adaptive pixel slimming and dynamic token slimming to preprocess and compress visual content.
DocLayLLM (2024)	Uses a document encoder to integrate image and positional info; employs CoT pretraining and annealing for high-quality synthetic data.
DocLLM (2024a)	Layout-aware transformer treating spatial text as a modality; uses disentangled self-attention and block masking for pretraining.
GPE (2025a)	Proposes Group Positional Embedding for novel positional encoding within MLLMs.
DocPedia (2024)	Uses frequency domain transformation for high-res visual encoding, improving visual and textual representation.
DoCo (2024)	Enhances LVLM visual representation using intra- and inter-document contrastive learning.
HRVDA (2024a)	Enables scalable high-res doc processing via visual token pruning and instruction/content filtering.
ICL-D3IE (2023)	In-context learning for inference, integrating labels, layout, and format via diverse demonstrations.
InstructDr (2024)	Bridges multimodal document and LLM representations with Document-former and multi-task instruction tuning.
LAPDoc (2024)	Introduces rule-based prompting for structured layout input (XML, HTML, Markdown, spatial format).
LayoutLLM (2024)	Layout-aware pretraining/fine-tuning with document encoder, LLM projection, and LayoutCoT for stepwise reasoning.
Leopard (2024)	Uses high-quality instruction-tuning data and high-res multi-image encoding.
LLaVA-Read (2024c)	Dual vision encoders extract fine/coarse-grained visual cues with diverse pretraining tasks.
LMDX (2024)	Designs prompting for KIE by encoding layout as a text prompt.
Marten (2025)	Introduces VQAMask for spatial-aware pretraining and synthetic mask vision-text alignment.
MDocAgent (2025)	Multi-agent framework integrating specialized agents for text/image understanding, extraction, and answer prediction.
mPLUG-DocOwl (2023a)	Multi-task instruction tuning (VQA, IE, NLI, IC), combining language-only and generation VL tasks.
mPLUG-DocOwl 1.5 (2024a)	Proposes H-Reducer for spatial-aware vision-to-text, shape-adaptive cropping, layout-aware instruction tuning.
mPLUG-DocOwl2 (2024b)	Visual token compression for high-res docs, shape-adaptive cropping, low-res encoder, multi-page tokens.
PDF-WuKong (2025)	Sparse sampling for relevant text/diagram extraction; contrastive learning applied.
PP-DocBee (2025)	Data synthesis strategy for QA with diverse-source sampling weights.
ProcTag (2025)	LLM-based process extraction, converts processes to pseudo-code for complexity/diversity analysis.
Texthawk (2024a)	ReSampling/ReArrangement for token compression; SPE, QPN, MLCA for hierarchical spatial encoding.
Texthawk2 (2024b)	16x token compression, unified encoder, detection head, multi-level cross-attention.
TextMonkey (2024c)	Addresses semantic continuity with Image/Token Resampler and position-aware interpretability tasks.
UReader (2023b)	Shape-adaptive cropping for high-res clarity; 2D positional encoding preserves layout.
Llavar (2024d)	Synthesizes large-scale instruction-tuning sets using templates and GPT-based generation.
TGDoc (2023)	Text-grounded instruction tuning for text and normalized bbox prediction.
UniDoc (2023)	Similar to TGDoc: text-grounded tuning with normalized bbox prediction.
KOSMOS-2.5 (2023)	Spatially-aware text block generation and structured markdown output in pretraining.
Vary (2024)	Scales up vision vocabulary for LVLMs via a vocabulary network merged with CLIP vocabulary.
LayTextLLM (2024)	Projects bounding boxes as single embeddings; interleaves with text; layout-aware and shuffled-OCR fine-tuning.
HVFA (2024)	Aggregates hierarchical visual features (feature pyramid); new tuning task for predicting relative positions.
TokenCorrCompressor (2024b)	Token-level correlation-guided sampling for efficient, parameter-free token selection.

Table 5: Key technical summaries of representative MLLM-based visually rich document understanding (VRDU) frameworks. Abbr.—KIE: Key Info Extraction, QA: Question Answering, IE: Information Extraction, VQA: Visual QA, NLI: Natural Language Inference, VL: Vision-Language

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

Paper	Set	Source	Size	Public Available
Vary	Document Data Engine	ArXiv, CC-MAIN, E-books	2M	No
	Chart Data Engine	matplotlib, pyecharts, NLP corpora	1.5M	No
	Detection Data Engine	Objects365, OpenImages	∼3M	Release Data Source
LLaVAR	LAION	LAION images filtered for text-rich content, OCR applied	0.4M	Yes
Doco	DoCo-Processed	CC3M (LLaVA) + LAION, processed with PaddleOCR	1.0M	No
Texthawk2	100M pretraining	Diverse, mainly public datasets	100M	No
Docpedia	PDF Images	arXiv (public scientific preprints)	325K	Yes
	PPT Images	Common Crawl (web-crawled PPTs)	600K	Partly

Table 7: 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	Yes/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 8: Summary of instruction-tuning datasets for Leopard and LLaVAR. *Partially: only part of the data is open.

Name	Conf./J.	Year	Domain	# Docs	# Images	# Keys / Qs	MP	Language	Metrics	Format	Task
FUNSD	ICDAR-w	2019	Multi-source	N/A	199	4	N	English	F1	P. & H.	KIE
SROIE	ICDAR-c	2019	Scanned Receipts	N/A	973	4	N	English	F1*	P.	KIE
CORD	NeurIPS-w	2019	Scanned Receipts	N/A	1,000	54	N	English	F1	P.	KIE
Payment-Invoice	ACL	2020	Invoice Form	N/A	14,832	7	N	English	F1	D.	KIE
Payment-Receipts	ACL	2020	Scanned Receipts	N/A	478	2	N	English	F1	P.	KIE
Kleister-NDA	ICDAR	2021	Private Agreements	540	3,229	4	Y	English	F1	D.	KIE
Kleister-Charity	ICDAR	2021	AFR	2,778	61,643	8	Y	English	F1	D. & P.	KIE
EPHOIE	AAAI	2021	Exam Paper	N/A	1,494	10	N	Chinese	F1	P. & H.	KIE
XFUND	ACL	2022	Synthetic Forms	N/A	1,393	4	N	Multilingual	F1	D. & P. & H.	KIE
Form-NLU	SIGIR	2023	Financial Form	N/A	857	12	N	English	F1	D. & P. & H.	KIE
VRDU-Regist. Form	KDD	2023	Registration Form	N/A	1,915	6	N	English	F1	D.	KIE
VRDU-Ad-buy Form	KDD	2023	Political Invoice Form	N/A	641	9+1(5)	N	English	F1	D. & P.	KIE
DocILE	ICDAR	2023	Invoice Form	6,680	106,680	55	Y	English	AP, CLEval	D. & P.	KIE
DocVQA	WACV	2021	Industrial Reports	N/A	12,767	50,000	N	English	ANLS	D./P./H.	VQA
VisualMRC	AAAI	2021	Website	N/A	10,197	30,562	N	English	BLEU, etc	D.	VQA
TAT-DQA	MM	2022	Financial Reports	2,758	3,067	16,558	Y	English	EM, F1	D.	VQA
RDVQA	MM	2022	Data Analysis Report	8,362	8,514	41,378	N	English	ANLS, ACC	D.	VQA
CS-DVQA	MM	2022	Industry Documents	N/A	600	1,000	N	English	ANLS	D./P./H.	VQA
PDFVQA-Task A	ECML-PKDD	2023	Academic Paper	N/A	12,337	81,085	N	English	F1	D.	VQA
PDFVQA-Task B	ECML-PKDD	2023	Academic Paper	N/A	12,337	53,872	N	English	F1	D.	VQA
PDFVQA-Task C	ECML-PKDD	2023	Academic Paper	1,147	12,337	5,653	Y	English	EM	D.	VQA
MPDocVQA	PR	2023	Industrial Reports	6,000	48,000	46,000	Y	English	ANLS	D./P./H.	VQA
DUDE	ICCV	2023	Cross-domain	5,019	28,709	41,541	Y	English	ANLS	D.	VQA
MMVQA	IJCAI	2024	Academic Paper	3,146	30,239	262,928	Y	English	EM, PM, MR	D.	VQA

Table 9: Benchmark datasets for Key Information Extraction (KIE) and Visual Question Answering (VQA) in visually rich documents.