Document Understanding and Multi-modal AI: A Survey

www.surveyx.cn

Abstract

This survey explores the transformative impact of multi-modal AI and document understanding technologies, emphasizing their critical role in enhancing information extraction accuracy and efficiency across diverse document types. The integration of large language models (LLMs) with visual cues has significantly advanced document understanding, particularly in form recognition tasks, as evidenced by improvements achieved through transformer models leveraging multi-modal inputs. Foundational datasets like SciAI and SciAD have been established, paving the way for future advancements. Notable methods such as TRIE demonstrate the benefits of end-to-end frameworks, outperforming state-of-the-art approaches in efficiency and accuracy. Additionally, the utilization of pre-trained models for subtasks has proven effective in enhancing multi-modal large language models (MLLMs), as demonstrated through extensive experimental results. Despite these advancements, challenges persist, particularly concerning dataset diversity, model calibration, and evaluation metrics, which are crucial for the applicability of multimodal document understanding models in enterprise contexts. The survey highlights the need for improved integration of visual cues in Key Information Extraction (KIE) systems and the potential of generative KIE methods. It underscores the importance of addressing data quality and model adaptability as critical areas for ongoing research. Furthermore, the potential of domain-specific frameworks, such as DAViD, to capture domain-specific knowledge using synthetic annotations is noted as a promising approach. In conclusion, while significant strides have been made, continued efforts are needed to refine these technologies, enhance dataset diversity, and improve model interpretability, facilitating broader adoption of Document AI technologies across various domains.

1 Introduction

1.1 Significance of Multi-modal AI

Multi-modal AI is crucial for advancing document understanding by integrating various data types, including textual and visual information, to enhance the interpretation of complex documents [1]. This integration is particularly vital for visually rich documents (VRDs), which present challenges due to their intricate structures and diverse visual cues. The evolution of large language models (LLMs) into multi-modal capabilities has further propelled this field, allowing for a nuanced understanding of document content through multiple modalities [2].

Incorporating both textual and visual elements addresses challenges in understanding visually situated language, significantly enhancing document comprehension [3]. This is exemplified in visually-situated text parsing (VsTP), where generative LLMs have advanced the automation of document understanding tasks [4]. Processing multimodal documents, which often include text, figures, and tables, necessitates sophisticated approaches for effective information extraction and question answering [5].

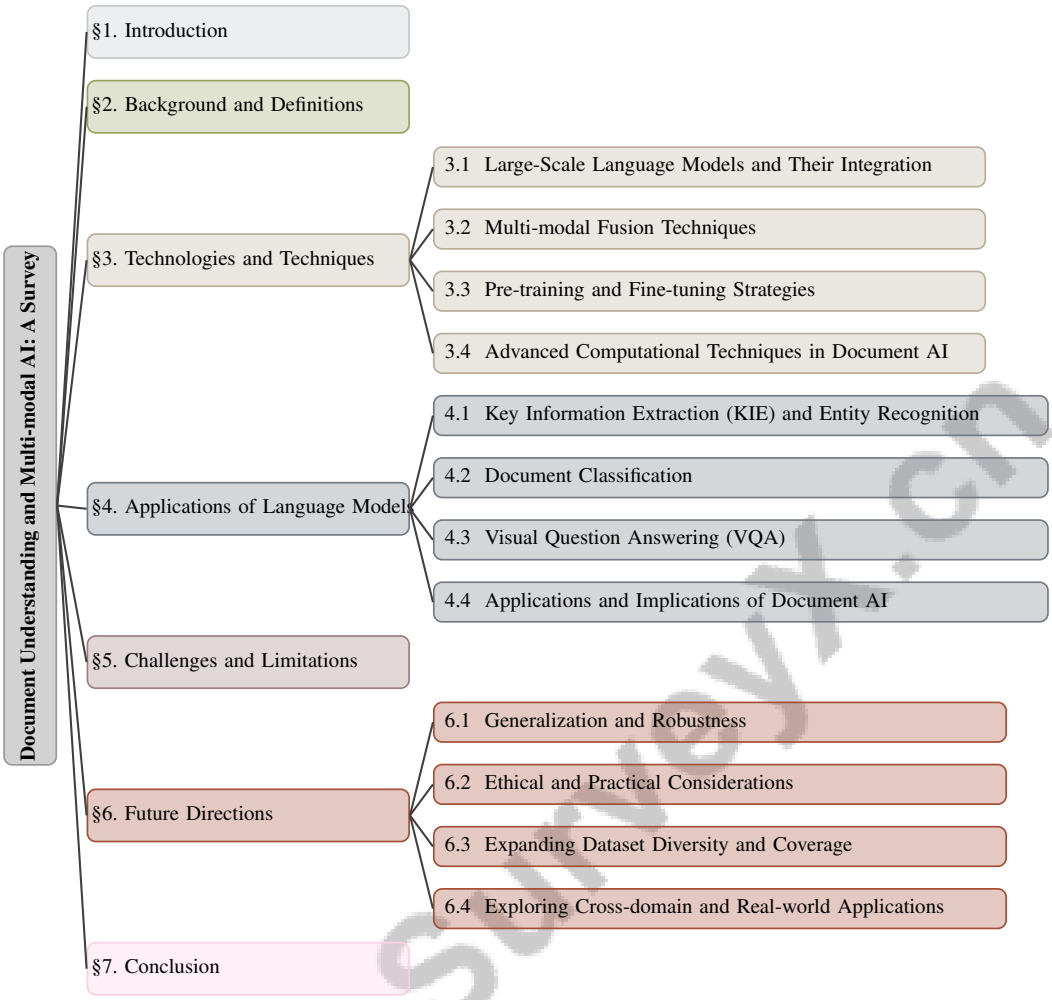


Figure 1: chapter structure

Despite advancements, challenges remain, such as aligning semantic information across modalities, particularly with closely resembling images [6]. Issues related to cognition and perception knowledge conflicts in MLLMs complicate document understanding, highlighting the need for consistent perception mechanisms [7]. Additionally, existing multimodal pre-training approaches struggle with markup-language-based documents, which require dynamic rendering of layout information [8].

The limitations of current LLMs and multimodal LLMs in analyzing scientific diagrams restrict their use in academic writing, indicating a need for future enhancement [9]. Privacy concerns in Document Visual Question Answering (DocVQA) models also pose significant challenges, particularly regarding the risks of memorization that could expose sensitive information [10].

As the demand for advanced automated document understanding solutions rises across industries, integrating multi-modal AI technologies is essential to address traditional methods' shortcomings. This integration enhances capabilities in document layout analysis, visual information extraction, and content classification, significantly improving the efficiency and accuracy of document AI systems. The evolution of deep learning techniques further facilitates processing complex document formats, contributing to successful digital transformation initiatives [11, 12, 13].

1.2 Objectives of the Survey

This survey aims to deliver an in-depth analysis of Document AI, focusing on the role of multi-modal large language models (MLLMs) in enhancing document comprehension across various fields. It explores advancements in text-grounding capabilities that improve the interpretation of textual

content within images, exemplified by the TGDoc model, which integrates spatial text positioning to bolster understanding in text-rich scenarios. The survey reviews evaluation methods for MLLMs, addressing applications across multiple domains and providing insights into performance metrics and benchmarks to foster the development of more effective document understanding technologies [14, 15, 16]. It also addresses limitations in existing NLP and computer vision methods, which often neglect the two-dimensional relationships between textual elements and visual features, impacting document segmentation, entity extraction, and attribute classification.

The primary objective is to assess the effectiveness of MLLMs in comprehension and generation tasks across modalities, particularly focusing on enhancing document AI capabilities through text-grounding techniques. This involves developing the TGDoc model, which improves the interpretation of text within images by recognizing spatial text positioning, achieving state-of-the-art results on multiple benchmarks [17, 15, 16, 18]. The survey also evaluates scalable document conversion services that maintain high throughput and responsiveness and advances data-model co-development for MLLMs by exploring data-centric approaches.

Additionally, the survey benchmarks OCR-free document understanding tasks to enable fair comparisons across models and provides a comprehensive overview of recent advancements in form understanding, particularly concerning scanned documents. It examines the relationship between document aesthetics and AI model prediction confidence, addressing the gap in understanding how aesthetic elements influence model behavior in document understanding tasks [7].

Novel approaches, such as leveraging distillation methods to harness the power of large LLMs while accommodating computational limitations, are introduced. The survey facilitates advancements in document understanding by providing comprehensive datasets for comparing and pretraining language models [9]. It proposes methods like DocVLM, which integrates OCR data into vision-language models to enhance document understanding while reducing computational overhead.

By systematically reviewing methodologies and challenges in multimodal document understanding, this survey aims to enhance the development of efficient models tailored to real-world applications, including extracting information from visually rich documents across various industries such as finance, law, and technology. Insights gained will serve as a foundation for researchers to address existing limitations in dataset curation, model effectiveness, and evaluation practices, ultimately facilitating the creation of adaptable models capable of operating effectively in diverse and resource-constrained environments [19, 20, 21, 13]. It highlights the potential for benchmarks to compare model performance in text-rich visual understanding tasks, providing a foundational framework for future research directions and practical applications in document AI.

1.3 Structure of the Survey

The survey is systematically organized to comprehensively explore document understanding and multi-modal AI, drawing inspiration from various structural frameworks in existing literature. It begins by introducing foundational concepts and the significance of multi-modal AI in document understanding, setting the stage for subsequent discussions. A detailed background section defines key terms and concepts essential for understanding the complexities of document AI, akin to the object-oriented approach in inefficient document parsing [22].

Core sections delve into the technologies and techniques underpinning document AI, categorized into sub-sections such as Large-Scale Language Models, Multi-modal Fusion Techniques, and Pre-training and Fine-tuning Strategies. This categorization mirrors the architectural and methodological segmentation found in existing research, facilitating a structured analysis of each component [23].

Subsequent sections explore the diverse applications of language models in document AI, highlighting specific tasks and successful implementations. This part emphasizes the interplay between different modalities and their contributions to enhancing document understanding, reflecting the synergy between data and MLLMs [24].

The survey also addresses challenges and limitations inherent in the field, providing a critical examination of issues such as data scarcity, model interpretability, and computational complexity. This section parallels discussions on the advantages and limitations of specific frameworks, offering insights into areas requiring further research and development [25].

Finally, the survey concludes with a forward-looking perspective on future directions, proposing strategies for improving model robustness, ethical considerations, and expanding applications across diverse domains. This concluding section synthesizes insights gained throughout the survey, reinforcing the foundational framework established in the introduction and background sections.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Document Understanding and Visually Rich Documents (VRDs)

Document understanding involves automating the extraction and interpretation of information from complex documents, particularly challenging in Visually Rich Documents (VRDs). These documents feature intricate layouts, diverse content types, and multi-page formats combining text, images, and spatial elements [26]. Traditional methods struggle with VRDs due to their reliance on serialized text representations, which miss crucial spatial relationships [27]. Large Visual Language Models (LVLMs) often fail to extract fine-grained features from text-rich images, leading to suboptimal performance in visually-situated natural language tasks [28].

The complexity of VRDs is heightened by varying visual styles and structural differences, posing challenges for Visual Question Answering (VQA) and Key Information Extraction (KIE) [29]. Traditional multi-stage methods, including Optical Character Recognition (OCR), demonstrate inefficiencies, highlighting the need for integrated solutions [30]. Additionally, the high computational demands of large pre-trained vision-language models limit their scalability across diverse document types [31].

Current benchmarks for document understanding often focus on limited pages and lack comprehensive layout analyses [3]. Challenges include processing numerical reasoning, understanding lengthy documents, and identifying interrelations among document elements [4]. Converting documents into machine-processable formats adds complexity due to format variability and intricate structures [32].

Advanced computational models integrating text, visual, and layout features are essential for enhancing VRD comprehension. This integration is crucial for overcoming existing methods' inefficiencies, which often fail to process complex document images effectively. Developing innovative solutions leveraging multiple modalities is vital for robust and scalable document AI solutions, fostering a deeper understanding of VRDs across applications [25]. Frameworks like DocMamba exemplify advancements by reducing computational complexity while retaining global modeling capabilities [31]. The PFL-DocVQA dataset, with its extensive question-answer pairs, serves as a valuable resource for evaluating and enhancing document understanding systems [10].

2.2 Multi-modal Large Language Models (MLLMs)

Multi-modal Large Language Models (MLLMs) advance Document AI by integrating diverse data modalities—text, images, and layout structures—into a cohesive analytical framework. This integration addresses traditional models' limitations, which often process each modality independently, enhancing document comprehension through a holistic approach [1]. GraphDoc, for instance, uses a multimodal graph attention-based model to learn contextualized document representations by combining textual, visual, and positional information [1].

A critical challenge for MLLMs is accurately perceiving and interpreting visual information, as current CLIP-like encoders often yield incomplete image representations. Enhanced semantic alignment techniques, such as Semantic Alignment for Multimodal large language models (SAM), incorporate bidirectional semantic guidance to improve visual data alignment before integration into the language model [9]. Benchmarks prompting LLMs to analyze scientific diagrams alongside textual context further emphasize integrating visual and textual modalities [9].

Models like DocFormerv2 integrate vision, language, and spatial features, enhancing document layout and context understanding using structural metadata and visual features from OCR. OCR-free models like Donut map document images to structured outputs without traditional OCR processes [1].

Innovative frameworks like DocMamba address transformers' limitations in Visually Rich Document Understanding (VrDU) by maintaining linear complexity while improving efficiency. Approaches

like MarkupLM, which jointly pre-train text and markup language, show potential for enhanced document understanding in markup-language contexts [9].

Despite advancements, MLLMs face conflicts between cognitive understanding and perceptual information, leading to inconsistent responses. For instance, in document VQA tasks, an MLLM may generate answers that do not align with visual content, revealing a disconnect between OCR perceptions and cognitive comprehension. This issue, termed Cognition and Perception (CP) knowledge conflicts, challenges MLLMs' performance and explainability. Findings indicate that even leading models like GPT-4o achieve only 68.6

By bridging data modalities, MLLMs enable a nuanced and holistic document analysis approach, paving the way for future innovations. Their proficiency in integrating diverse information sources highlights their essential contribution to Document AI technologies' evolution. This advancement facilitates sophisticated and scalable solutions for complex document understanding tasks, addressing challenges like varied layouts, low-quality scanned images, and intricate business document structures. Leveraging deep learning techniques for tasks like document layout analysis and visual information extraction enhances productivity and supports ongoing digital transformation across industries [11, 33, 34, 12].

2.3 Natural Language Processing (NLP) in Document AI

Natural Language Processing (NLP) is crucial in Document AI, facilitating textual information extraction and interpretation from documents and enhancing complex document structure understanding. Key Information Extraction (KIE) is fundamental, focusing on identifying and extracting pertinent data from document images. Integrating contextual and spatial semantics in a two-dimensional (2D) space is vital for accurately capturing relationships between textual elements, as evidenced by methods incorporating these aspects to improve KIE performance [35].

Document AI employs various NLP techniques to address tasks such as Document Classification (CLS), Page Segmentation (SEG), and Visual Question Answering (VQA). These tasks are organized into a coherent framework that enhances the efficient processing and understanding of diverse document types. For instance, page segmentation techniques analyze and divide documents into logical sections, a critical step for subsequent information extraction and classification [19].

The processing efficiency of high-resolution document images poses a significant challenge for MLLMs, as the large number of visual tokens generated can hinder scalability and performance. Innovative approaches focusing on token-level correlation-guided compression aim to reduce visual tokens without compromising information quality, enhancing overall document processing efficiency [36].

NLP tools and methods are continuously evolving, offering advanced capabilities for document AI applications, including sophisticated algorithms for tabular reasoning that enable data extraction and interpretation from structured tables within documents. Furthermore, visual question answering systems are refined to better manage complex queries and provide accurate responses based on documents' visual and textual content [19].

By integrating NLP techniques with multi-modal data processing, Document AI systems achieve a more comprehensive understanding of documents, facilitating improved information extraction and decision-making across various applications. This integration underscores NLP's critical role in enhancing document AI technologies, essential for automating the reading, understanding, and analysis of complex business documents. Leveraging deep learning advancements and various model architectures, including transformer-based and graph-based approaches, fosters robust and scalable solutions for diverse document understanding tasks, such as layout analysis, information extraction, and visual question answering. Ultimately, these advancements enhance productivity and efficiency in digital transformation efforts and address challenges posed by the variety of document formats and layouts encountered in real-world applications [11, 33, 12, 37].

3 Technologies and Techniques

The exploration of emerging technologies and techniques in document AI reveals a transformative landscape characterized by the integration of advanced methodologies and innovative frameworks.

This section delves into significant advancements that have reshaped the field, particularly focusing on large-scale language models (LLMs) and their integration with multi-modal data, which enhance document understanding and performance in document AI systems. Table 1 provides a comprehensive comparison of various methodologies and approaches in document AI, focusing on their integration techniques, data modalities, and efficiency objectives.

Figure 2 illustrates the hierarchical structure of these emerging technologies and techniques in Document AI, highlighting the integration of Large-Scale Language Models, Multi-modal Fusion Techniques, Pre-training and Fine-tuning Strategies, and Advanced Computational Techniques. Each category is further divided into specific applications, frameworks, and methods that contribute to enhanced document understanding, processing efficiency, and adaptability across diverse document formats. This visual representation not only underscores the complexity of the field but also emphasizes the interconnectedness of various methodologies that drive innovation in document AI.

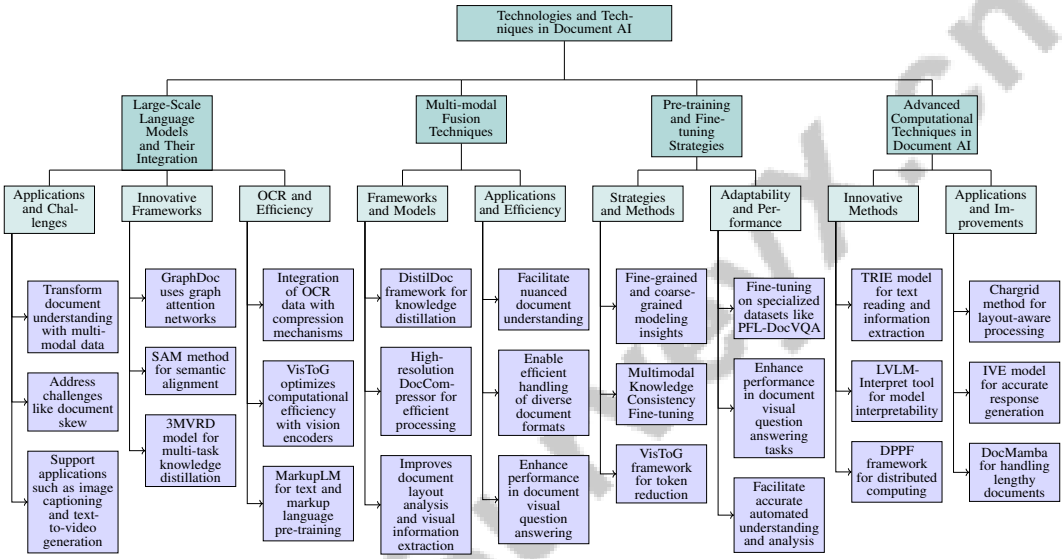


Figure 2: This figure illustrates the hierarchical structure of emerging technologies and techniques in Document AI, highlighting the integration of Large-Scale Language Models, Multi-modal Fusion Techniques, Pre-training and Fine-tuning Strategies, and Advanced Computational Techniques. Each category is further divided into specific applications, frameworks, and methods that contribute to enhanced document understanding, processing efficiency, and adaptability across diverse document formats.

3.1 Large-Scale Language Models and Their Integration

Large-scale multi-modal language models (MMLLMs) fundamentally transform document understanding through the integration of diverse data types—text, images, and audio—improving information extraction and interpretation from complex documents. Studies underscore their effectiveness in addressing challenges like document skew, which affects extraction accuracy, advocating for robust architectures to enhance real-world performance. This evolution supports advanced applications such as image captioning and text-to-video generation, emphasizing the importance of responsible AI practices [38, 39].

Innovative frameworks like GraphDoc use graph attention networks to contextualize information from various semantic regions, enhancing multi-modal data integration [1]. The SAM method addresses semantic misalignment in MLLMs by incorporating contextual semantics from images during visual token extraction, improving visual data alignment before integration into language models [6]. The 3MVRD model exemplifies integrating multi-task and multi-teacher joint-grained knowledge distillation for understanding visually-rich form documents [40].

Integrating OCR data is critical, as shown by models incorporating OCR encoders with compression mechanisms to distill OCR data into learned queries, enhancing document understanding while reducing computational overhead. VisToG uses pre-trained vision encoders to cluster similar image

segments into semantically related concepts, optimizing computational efficiency by reducing the number of visual tokens processed by the LLM [41].

Frameworks like MarkupLM integrate text and markup language pre-training within a single framework, enhancing the understanding of markup-based documents [8]. Additionally, Fox facilitates a pipeline enabling LVLMs to focus on specific regions of interest in single or multi-page documents, improving their ability to perform fine-grained tasks like OCR and translation [28].

Benchmarking efforts, including assessments of proprietary models like GPT-4o and open-source models like Qwen2-VL-7B, are essential for evaluating LLM performance on diverse tasks, highlighting the necessity for continuous innovation in model development [5]. Experiments with mPLUG-DocOwl further demonstrate the capabilities of state-of-the-art multimodal LLMs as baseline models [9].

By employing advanced end-to-end models and techniques, researchers can significantly enhance the automation of reading, understanding, and analyzing complex documents—such as invoices, tickets, and resumes. This integrated approach fosters deeper comprehension of the rich information contained within these documents, leading to improved information extraction and decision-making processes across various applications, including document layout analysis and visual information extraction [11, 42, 34, 12]. This ultimately facilitates more robust and scalable solutions for document AI technologies.

3.2 Multi-modal Fusion Techniques

Multi-modal fusion techniques are pivotal for enhancing document AI performance by effectively combining diverse data modalities, such as text, images, and layout structures. These techniques create cohesive representations that leverage the strengths of each modality, improving comprehension and interpretation of complex documents. The DistilDoc framework focuses on distilling knowledge from complex teacher models to simpler student models, enhancing layout awareness in document processing while preserving essential layout features [43].

The High-resolution DocCompressor compresses high-resolution document images into a fixed number of visual tokens, maintaining both layout and textual information. This architecture improves processing efficiency without compromising the quality of extracted information, especially in scenarios requiring rapid processing of high-resolution images [44].

These fusion techniques maximize the strengths of each data modality, facilitating nuanced and comprehensive document understanding. By employing advanced techniques such as knowledge distillation and sophisticated compression mechanisms, Document AI systems enhance their performance in synthesizing multi-modal information, improving effectiveness in tasks like document layout analysis, visual information extraction, and document visual question answering. These strategies enable efficient handling of diverse document formats and layouts while maintaining high accuracy in understanding complex business documents [33, 11, 43, 37, 12]. This ultimately leads to more robust and scalable solutions for document understanding.

As shown in Figure 3, this figure illustrates key multi-modal fusion techniques in document AI, highlighting the DistilDoc framework for knowledge distillation and layout awareness, the High-resolution DocCompressor for efficient visual token compression, and various applications such as layout analysis and visual question answering. The advent of multi-modal fusion techniques has significantly enhanced AI systems' capabilities by enabling the integration of information from multiple sources and modalities. The "Online and Off-line Document Classification Framework" combines pre-processed (off-line) and real-time (on-line) stages to classify documents, utilizing a similarity measure for feature comparison and generating new classes with limited samples. The "Multi-modal Image Retrieval with a Novel Entity-Based Retrieval Model" integrates text, image, and multi-modal queries to improve search and retrieval, particularly in identifying specific styles through structured queries. Lastly, the "Large-scale Multi-modal Dataset" flowchart illustrates advanced architectures and techniques in multi-modal learning and diffusion, highlighting the integration of vision, language, and video into a unified framework. Collectively, these examples demonstrate the diverse applications and transformative potential of multi-modal fusion techniques in machine learning and AI [45, 46, 47].

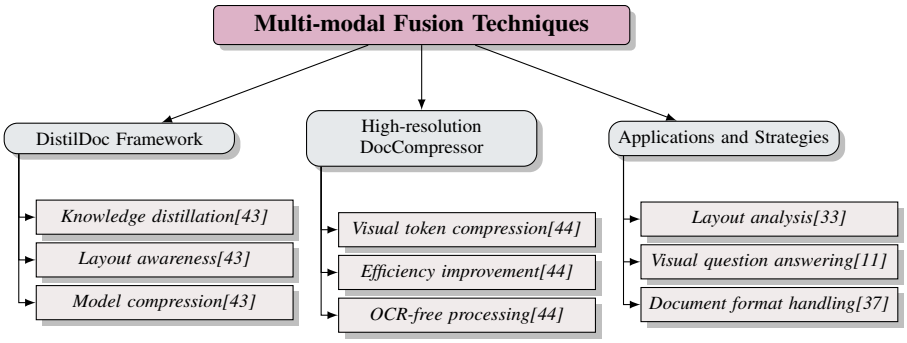


Figure 3: This figure illustrates key multi-modal fusion techniques in document AI, highlighting the DistilDoc framework for knowledge distillation and layout awareness, the High-resolution DocCompressor for efficient visual token compression, and various applications such as layout analysis and visual question answering.

3.3 Pre-training and Fine-tuning Strategies

Pre-training and fine-tuning strategies are crucial for optimizing model performance in document AI tasks by leveraging diverse data modalities. The integration of fine-grained and coarse-grained modeling insights, exemplified by the 3MVRD method, enhances understanding of form documents, bridging gaps between different granularity levels [40].

The Multimodal Knowledge Consistency Fine-tuning method improves overall model consistency through specific tasks: Cognition Consistency, Perception Consistency, and CP Connector tasks, emphasizing the maintenance of consistency between cognitive and perceptual knowledge to enhance model reliability [7].

The VisToG framework employs a novel grouping mechanism that leverages pre-trained vision encoders to group similar image segments into semantically meaningful tokens, significantly reducing the number of processed tokens while maintaining performance [41].

Fine-tuning on specialized datasets, such as the PFL-DocVQA, under various configurations—including centralized and federated learning—demonstrates model adaptability to different learning environments, which is crucial for enhancing performance in document visual question answering tasks and ensuring robust, privacy-aware solutions [10].

These strategies underscore the importance of tailored pre-training and fine-tuning processes in advancing document AI technologies, enabling the development of robust models capable of effectively understanding and processing complex documents across various domains. By integrating advanced strategies such as deep learning techniques for document layout analysis, visual information extraction, and active learning frameworks like OPAD, document AI systems can enhance performance and scalability. This progress facilitates accurate automated understanding and analysis of diverse document formats, laying the groundwork for future innovations that address document structure complexities and improve productivity in business processes [11, 33, 12, 25].

3.4 Advanced Computational Techniques in Document AI

Advanced computational techniques have significantly enhanced Document AI capabilities through innovative methods that improve the integration and understanding of complex documents. The TRIE model exemplifies this advancement by leveraging the mutual influence between text reading and information extraction, enhancing document content understanding [42].

The LVLM-Interpret tool offers functionalities for visualizing raw attentions, relevancy maps, and causal interpretations, providing insights into how model outputs relate to input images [48]. This tool enhances the interpretability of large vision-language models, allowing a deeper understanding of decision-making processes in document AI systems.

DPPF, a distributed computing framework, facilitates parallel processing of large datasets through interconnected nodes, improving scalability and efficiency in handling extensive data volumes [49].

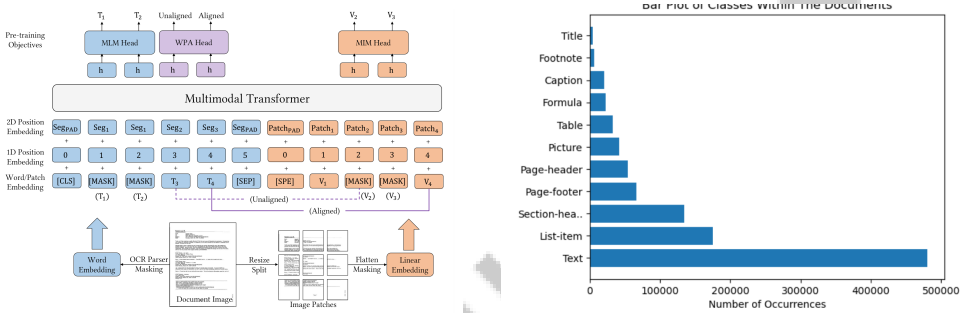
The chargrid method preserves spatial relationships between characters, enabling accurate segmentation and classification of text based on layout, which is crucial for layout-aware processing [50].

Innovations like the IVE model, which integrates diverse visual information sources, enhance the model's ability to generate accurate responses to visually complex documents [51].

DocMamba captures long-range dependencies with linear-time inference, allowing it to handle lengthy documents without losing contextual information, which is vital for maintaining coherent understanding [31].

Fox leverages position-aware prompts and hybrid visual vocabularies for fine-grained document understanding, demonstrating the effectiveness of integrating spatial and semantic cues [28].

Recent advancements in computational methods for Document AI lead to the development of more robust, efficient, and interpretable solutions for automated reading, understanding, and processing of complex documents. These improvements in document layout analysis, visual information extraction, and classification, driven by deep learning technologies, are crucial for managing diverse formats and structures of digital and scanned documents, ultimately facilitating faster, more accurate information processing across various applications and supporting ongoing digital transformation in industries [52, 11, 33, 12].



(a) Multimodal Transformer for Document Image Segmentation and Recognition[53]

(b) Bar Plot of Classes within The Documents[33]

Figure 4: Examples of Advanced Computational Techniques in Document AI

As shown in Figure 4, advanced computational techniques are pivotal in enhancing the efficiency and accuracy of document processing. The first image illustrates a multimodal transformer architecture designed for document image segmentation and recognition, integrating a pre-training objective module with specialized heads for learning document image representation and aligning it. This design emphasizes the importance of multimodal approaches in capturing complex document structures. The second image presents a bar plot quantifying occurrences of various document classes, such as Title, Footnote, and Table, providing insights into their distribution and aiding in refining document processing strategies. Together, these examples highlight significant advancements in computational techniques driving innovations in Document AI [53, 33].

Feature	Large-Scale Language Models and Their Integration	Multi-modal Fusion Techniques	Pre-training and Fine-tuning Strategies
Integration Approach	Multi-modal Fusion	Knowledge Distillation	Granularity Modeling
Data Modality	Text, Images, Audio	Text, Images, Layout	Form Documents
Efficiency Focus	Real-world Performance	Processing Efficiency	Model Adaptability

Table 1: This table presents a comparative analysis of integration approaches, data modalities, and efficiency focuses across three key areas in document AI: Large-Scale Language Models and Their Integration, Multi-modal Fusion Techniques, and Pre-training and Fine-tuning Strategies. It highlights the distinct methodologies and data types utilized in each category, emphasizing their respective contributions to enhancing document understanding and processing efficiency.

4 Applications of Language Models

4.1 Key Information Extraction (KIE) and Entity Recognition

Key Information Extraction (KIE) and Entity Recognition are pivotal in Document AI, emphasizing the accurate extraction and identification of entities from documents with intricate layouts. The efficacy of these processes is significantly enhanced by advanced models that incorporate multimodal data, facilitating the precise extraction of key-value pairs. For instance, the HRVDA model showcases the ability to process high-resolution images directly, thereby improving accuracy in information extraction and visual question answering tasks [54].

Integrating visual features with textual data is essential, as evidenced by the OCR-free document understanding transformer, which markedly boosts KIE performance [30]. Similarly, frameworks like OmniParser achieve state-of-the-art performance across various tasks, demonstrating their effectiveness in parsing visually situated text [4].

KIE model evaluations often utilize datasets such as FUNSD, CORD, and SROIE for information extraction, alongside RVL-CDIP for document classification, providing robust platforms for performance assessment [55]. Metrics like precision, recall, and F1 scores are critical for evaluating KIE models, as illustrated in tasks including section splitting and entity extraction [32].

Advanced techniques, such as those in the DavarOCR toolbox, which encompasses text detection, recognition, KIE, and layout analysis, underscore the importance of integrating OCR capabilities to enhance KIE [29]. Moreover, models like 3MVRD, evaluated on datasets like FUNSD and FormNLU, highlight the significance of modality-aware approaches in boosting entity extraction accuracy [40].

Through the application of these advanced techniques and diverse datasets, KIE and Entity Recognition are evolving, enabling more precise and efficient information extraction from complex documents. This evolution is crucial for enhancing Document AI technologies, which leverage cutting-edge natural language processing and computer vision to automatically read, understand, and analyze various document formats. Improvements in tasks such as document layout analysis, visual information extraction, and document classification facilitate better decision-making and streamline information retrieval across numerous applications, ultimately driving productivity and supporting digital transformation initiatives [11, 33, 12].

4.2 Document Classification

Document classification is a fundamental task in Document AI, focusing on categorizing documents based on content and structural features, which is essential for efficient information organization and retrieval. Recent advancements in language models have significantly improved the accuracy and robustness of document classification systems. For instance, the LiLT model has proven effective in document classification tasks, achieving high F1 scores across multiple languages, thereby underscoring its language-independent capabilities [56].

The StructuralLM model exemplifies the integration of structural information into language models, enhancing document classification by utilizing layout features alongside textual content. Fine-tuned on various downstream tasks, including form understanding and document classification, this model demonstrates versatility in handling complex document structures [57].

End-to-end frameworks like Dessurt have been evaluated across diverse document datasets, encompassing tasks such as document classification, question answering, form understanding, and handwriting recognition. These evaluations highlight the model's ability to generalize across different document types and tasks, providing a comprehensive solution for document understanding [58].

Furthermore, the exploration of zero-shot and few-shot learning paradigms in document classification reveals that current models can generalize to new tasks with minimal training data. This capability is vital for adapting to novel document types and applications, achieving promising accuracy levels even with limited labeled data [59].

By leveraging these advanced models and techniques, document classification systems can enhance performance, facilitating efficient information retrieval and decision-making across various domains. This progress is crucial for significantly improving Document AI technologies, which encompass automated reading, understanding, and analysis of complex business documents. Utilizing deep

learning techniques for document layout analysis and visual information extraction not only enhances current applications but also lays the groundwork for future innovations in natural language processing and computer vision, ultimately driving the success of digital transformation initiatives across industries [11, 33, 12].

4.3 Visual Question Answering (VQA)

Visual Question Answering (VQA) is a crucial component of Document AI that enables the interpretation of complex documents by integrating visual and textual data to answer queries accurately. This task requires the fusion of multimodal information, exemplified by the TAT-DQA dataset, which consists of real-world financial documents necessitating both extraction and arithmetic reasoning across multi-page formats [60]. The complexity of VQA tasks is further illustrated by SlideVQA, which introduces multi-hop reasoning over multiple images, including annotated arithmetic expressions for numerical reasoning, thereby enhancing visual document understanding [61].

The TGDoc framework emphasizes focusing on relevant textual regions within text-rich images, significantly improving performance in visual question answering tasks [16]. This focus on pertinent regions is essential for enhancing comprehension and accuracy in document interpretation. Similarly, the model evaluated by Cao et al. highlights the implications of attention mechanisms in understanding complex documents, underscoring the significance of precise visual and textual alignment in VQA [52].

Advanced models such as IVE have been assessed using various multimodal datasets, including VQA and OCR-related datasets, demonstrating superior performance against state-of-the-art methods [51]. Evaluations conducted by Kim et al. on diverse Document VQA benchmarks further showcase the competitive performance of models like Cream in visually situated natural language understanding tasks [62].

Benchmark studies by Lu et al. incorporate datasets like TDIUC, TallyQA, and DVQA, designed to test different aspects of visual reasoning and understanding, providing a comprehensive evaluation framework for VQA systems [18]. Additionally, Liu et al. emphasize the evaluation of fine-grained sub-tasks, including region-level OCR and multi-page VQA, using a diverse set of documents, highlighting the need for precision in multi-page document analysis [28].

By leveraging advanced models and datasets, VQA in Document AI continues to evolve, enabling more accurate and efficient understanding of complex documents. This advancement is vital for enhancing Document AI technologies, which focus on automated reading, understanding, and analysis of diverse business documents. The integration of deep learning techniques for tasks such as document layout analysis, visual information extraction, and document classification not only improves current functionalities but also sets the stage for future innovations in natural language processing and computer vision, ultimately driving the success of digital transformation initiatives across industries [63, 33, 11, 64, 12].

4.4 Applications and Implications of Document AI

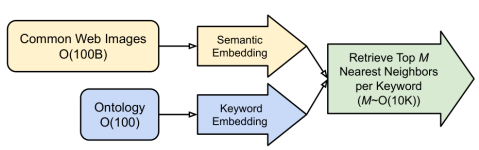
Document AI technologies exhibit transformative potential across various domains, enhancing information extraction and understanding from complex documents. The integration of multi-modal large language models (MLLMs) has advanced Visual Document Understanding (VDU), with models like DocFormerv2 achieving state-of-the-art performance across multiple VDU tasks, emphasizing the importance of local feature alignment and novel pre-training tasks in enhancing model comprehension [27]. Similarly, DocVLM improves document understanding capabilities in vision-language models (VLMs), achieving state-of-the-art results on benchmarks while maintaining computational efficiency, underscoring its practical applications [65].

In practical applications, UDOP has achieved state-of-the-art results on eight document AI tasks, demonstrating its effectiveness across diverse contexts [26]. The SAM model significantly outperforms existing methods, showcasing superior semantic alignment and coherence in multi-modal instructions, effectively addressing challenges associated with diverse image contexts [6]. Additionally, Rationale Distillation has been shown to enhance the performance of small image-to-text models on visual document understanding tasks, achieving higher accuracy with minimal computational cost [3].

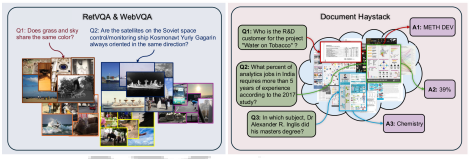
The application of document AI extends to scientific domains, where comprehensive datasets support the analysis of multiple diagrams in academic writing, contributing to the development of a mathematical knowledge base [9]. MarkupLM, integrating text and markup language pre-training, significantly outperforms existing models in document understanding tasks, demonstrating its effectiveness in practical applications [8].

To improve the truthfulness and ethical alignment of LLMs, visual instruction tuning has shown substantial improvements, surpassing traditional methods like Reinforcement Learning from Human Feedback (RLHF) [2]. GraphDoc achieves state-of-the-art performance in various document understanding tasks, demonstrating successful applications in form understanding and document classification [1].

The case studies and examples presented illustrate the effective deployment of Document AI technologies across a variety of applications, showcasing their capability to significantly enhance information extraction, improve document comprehension, and streamline decision-making processes in multiple sectors. This advancement is driven by deep learning techniques that facilitate complex tasks such as document layout analysis, visual information extraction, and document classification, ultimately supporting the digital transformation of businesses by enabling automated and accurate processing of diverse document formats [11, 33, 12, 20]. By leveraging advanced models and techniques, Document AI continues to evolve, paving the way for future innovations and broader applicability in the field.



(a) Semantic Embedding and Keyword Embedding for Image Retrieval[66]



(b) RetVQA WebVQA: A Comparative Study of Question Answering Techniques[67]

Figure 5: Examples of Applications and Implications of Document AI

As illustrated in Figure 5, the examples provided delve into the diverse applications and implications of Document AI, highlighting two specific instances: "Semantic Embedding and Keyword Embedding for Image Retrieval" and "RetVQA WebVQA: A Comparative Study of Question Answering Techniques." The first example presents a flowchart outlining the process of semantic and keyword embedding for effective image retrieval, beginning with a vast dataset of common web images transformed into semantic embeddings. These embeddings are then utilized to identify the top nearest neighbors per keyword, demonstrating Document AI's potential in enhancing image retrieval systems. The second example offers a comparative study of two question answering techniques, RetVQA and WebVQA, through a visual representation that includes a document haystack and sample questions, exploring their capabilities in answering complex queries, such as determining color similarities between grass and sky or identifying details about historical Soviet space missions. Together, these examples underscore the significant advancements and possibilities in Document AI, showcasing its ability to transform data interaction and retrieval processes [66, 67].

5 Challenges and Limitations

Document AI systems face significant challenges, primarily revolving around data quality and scarcity, model interpretability, and computational efficiency. These issues are critical as they influence the performance and applicability of AI models in real-world scenarios.

5.1 Data Scarcity and Quality

The development of Document AI models is critically hindered by data scarcity and quality, which affect their generalizability across diverse applications. The lack of comprehensive datasets that reflect the complexity of real-world documents is a significant issue, compounded by the dependency on large annotated datasets for model training, limiting scalability across domains [7, 40]. High-quality data is essential; low-resolution images and poor OCR outputs can severely degrade model performance [28, 27]. Additionally, models often struggle with non-English documents and complex

layouts due to inadequate character embeddings [26]. Real-world document processing, especially multi-page documents, remains challenging due to current models' focus on single-page processing [28]. The computational demands of using OCR and LLMs during inference further complicate efficiency [3]. Privacy concerns in training DocVQA models and limitations in handling diverse document types also present significant hurdles [10, 8]. To advance Document AI, enhancing data collection, annotation practices, and developing diverse datasets are crucial. Initiatives like DUDE are pivotal, introducing benchmarks that address real-world challenges, aiding the automation of business document understanding and analysis [20, 11].

5.2 Model Interpretability and Transparency

Model interpretability and transparency are critical in the deployment of Document AI systems, especially for MLLMs that integrate diverse data modalities. The complexity of these models often obscures decision-making processes, posing risks in high-stakes applications [68]. Misinterpretations and biases within MLLMs necessitate robust explanations to enhance transparency [69]. Attention values, commonly used for interpretation, often fail to capture causal relationships, leading to oversimplified explanations [48]. The lack of transparency in training data and methodologies poses challenges for reproducibility and ethical compliance [39]. To improve interpretability, sophisticated tools and methodologies are needed, particularly for complex enterprise documents. These tools must enhance transparency while upholding ethical standards [19, 11].

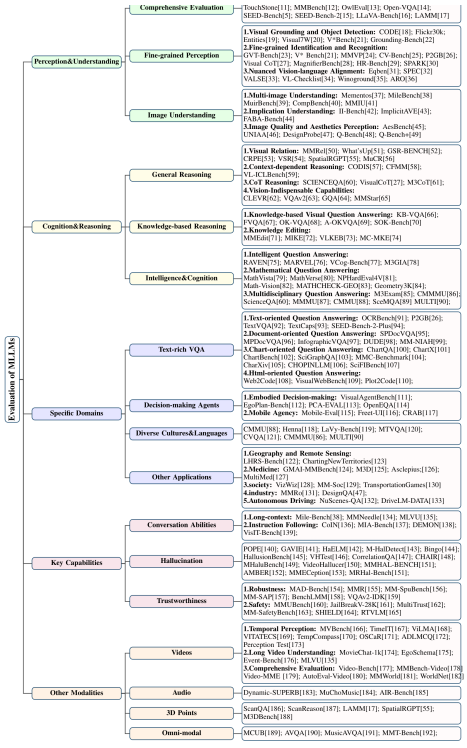
5.3 Computational Complexity and Efficiency

Document AI solutions are often limited by computational complexity and efficiency challenges, particularly due to the high demands of processing extensive data volumes. The quadratic computational and memory requirements of Transformer models' self-attention mechanisms hinder scaling to longer documents [70]. Image token redundancy further adds to computational overhead, limiting MLLMs' deployment in resource-constrained environments [41]. Methods like LiLT offer potential solutions by efficiently transferring knowledge to multilingual tasks, but substantial resources are still needed for pre-training on large datasets [71]. Balancing resolution needs for accuracy with computational efficiency remains a critical concern, particularly for models requiring high-resolution inputs [65]. Strategies enhancing computational efficiency while maintaining accuracy are essential, leveraging deep learning for tasks like layout analysis and content classification. Active learning frameworks like OPAD optimize annotation processes, improving performance in content detection tasks [33, 19].

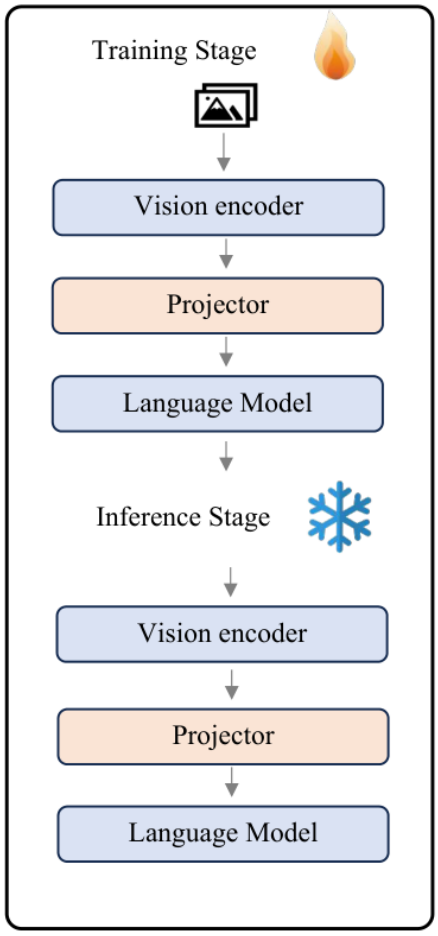
5.4 Advanced Techniques and Model Interpretability

Balancing advanced techniques with model interpretability is crucial in Document AI, as sophisticated models integrate multiple modalities and complex architectures. The opacity of multimodal document understanding models challenges real-world applications, particularly in enterprise settings where transparency is essential [19, 72]. To enhance interpretability, innovative methodologies are being explored, focusing on transparency and explainability without compromising performance. Techniques such as attention visualization and relevance mapping aim to clarify how models process data [68]. Establishing interpretability frameworks that address challenges posed by multimodal data is essential for enhancing trust in AI systems, particularly in high-stakes environments [73, 74].

As shown in Figure 6, understanding the complexities and limitations of advanced techniques in machine learning is crucial for developing effective and interpretable models. The figure illustrates these challenges, with a tree diagram categorizing evaluation metrics and tasks, emphasizing a comprehensive framework encompassing perception, reasoning, cognition, and intelligence. The flowchart details a machine learning model's process, highlighting the vision encoder and projector's role in data processing, underscoring the interconnected nature of these processes. These illustrations provide insight into sophisticated techniques and the ongoing quest for greater model interpretability despite inherent challenges [75, 68].



(a) The image is a tree diagram illustrating the evaluation of machine learning models (MLMs) across various domains and modalities.[75]



(b) A Flowchart of a Machine Learning Model[68]

Figure 6: Examples of Advanced Techniques and Model Interpretability

6 Future Directions

6.1 Generalization and Robustness

Enhancing the generalization and robustness of Document AI models is crucial for their effective deployment across diverse document types and contexts. Future research should focus on refining model capabilities through improved OCR integration and resilience against varying document qualities [30]. Expanding datasets to encompass a broader range of document formats, as seen in initiatives like M6Doc, can bolster model robustness and address challenges posed by uncommon layouts [40]. Synthetic data generation techniques and domain adaptation strategies, such as the DAViD framework, can further augment model accuracy by simulating diverse document scenarios [40].

Research should also focus on enhancing adaptability to non-Latin writing systems and investigate self-supervised learning objectives to improve generalization across linguistic contexts [26]. The exploration of unified interactive multimodal large language models (MLLMs) that incorporate visual experts may further enhance capabilities [51]. Innovative methodologies for visual instruction tuning and new MLLM architectures present promising avenues for improving model alignment and robustness [2].

Standardizing prompt engineering methods and expanding benchmarks to cover additional visual reasoning tasks will provide a comprehensive evaluation framework for model performance [18]. Pursuing these directions can lead to the development of robust, generalizable Document AI models capable of delivering accurate performance across various applications [4]. Additionally, enhancing visual vocabularies and addressing document-level tasks can improve the capabilities of large vision-language models (LVLMs) [28]. Refining contextual semantics extraction processes and diversifying datasets will further contribute to the robustness of Document AI models [6]. Moreover, advancing privacy-preserving techniques and developing datasets for various document types are crucial for enhancing model resilience [10].

6.2 Ethical and Practical Considerations

The deployment of Document AI technologies necessitates careful examination of ethical implications and practical considerations, especially as these systems integrate into critical applications. Multimodal large language models (MLLMs) pose significant ethical challenges, particularly regarding the perpetuation of biases and the generation of hallucinations, which are concerning in high-stakes domains where interpretability is crucial for reliable decision-making [69].

Ethical implications also arise in model development and deployment, where the choice between proprietary and open-source models raises questions of transparency, accountability, and accessibility. Proprietary models often limit adaptability, hindering innovation and equitable access to AI technologies [39]. In contrast, open-source models offer transparency but may lack the support available to proprietary solutions, creating a critical balance in MLLM deployment.

Practical considerations include the significant computational costs associated with developing and maintaining advanced models. The resource-intensive nature of training and fine-tuning multimodal models presents scalability challenges, particularly for organizations with limited resources [39]. Innovative approaches to model optimization and resource allocation are essential for making Document AI accessible to a broader range of users.

Additionally, ethical concerns related to data biases and privacy are paramount during model development. Ensuring training datasets are representative and free from biases is crucial for developing fair AI systems [62]. Privacy considerations are equally critical, as the use of sensitive data can lead to breaches and misuse. Implementing robust privacy-preserving techniques and complying with data protection regulations are essential for mitigating these risks [47].

6.3 Expanding Dataset Diversity and Coverage

Expanding dataset diversity and coverage is pivotal for advancing Document AI technologies, enabling effective generalization across various document types. Future research should focus on including additional document types and refining synthetic generation pipelines to enhance realism and complexity representation [76]. This approach ensures models are exposed to a broader spectrum of document structures, ultimately improving adaptability and performance.

Integrating visual capabilities into existing models, such as LLaSM, by combining vision and audio modalities can significantly enhance multi-modal interaction and dataset diversity [77]. This integration facilitates a comprehensive understanding of documents by leveraging multiple sensory inputs, enriching datasets with diverse modalities.

Enhancing image quality through advanced super-resolution techniques and integrating high-accuracy outputs from sophisticated Optical Character Recognition (OCR) systems can elevate model output precision in document understanding. This improvement is crucial in fields like law, finance, and technology, where accurate interpretation of complex layouts is essential. Recent advancements demonstrate the superior performance of models like GPT-4 Vision Turbo when provided with both OCR-recognized text and document images. Moreover, understanding aesthetic aspects of document design, such as layout quality and legibility, can further enhance model confidence and performance, emphasizing the need for comprehensive approaches that consider both visual and textual data in document analysis [63, 78, 21, 72].

To effectively address the unique challenges posed by diverse document structures, integrating various document types and evaluating their influence on model performance across languages and domains is essential. Research highlights significant variations in content and structure across

disciplines, underscoring the importance of pragmatic analysis alongside semantic understanding. The development of diverse prompt-response datasets, such as K2Q, illustrates that employing varied question formats enhances model robustness, particularly in visually rich document understanding tasks [79, 80]. By expanding datasets to include a broader range of formats, researchers can develop models that robustly address the nuances of different document types.

Exploring new model architectures that leverage the unique challenges of expanded benchmarks will further advance Document AI capabilities. Refining existing models to analyze non-Manhattan document layouts is a significant challenge in document layout analysis (DLA). This involves developing an image layer modeling method and introducing the FPD dataset, the first manually labeled fine-grained segmentation dataset for non-Manhattan layouts. Additionally, integrating aesthetic guidance into the synthesis process—drawing on principles of document design that emphasize visual elements like layout quality, font contrast, and alignment—can improve the realism and applicability of synthetic datasets. This comprehensive approach enhances the extraction of fine-grained features from complex document structures, bolstering overall AI model performance in visual document understanding tasks [63, 81, 72].

Implementing these strategies will enhance the development of resilient, flexible, and precise AI solutions for document understanding. This advancement will improve various applications—such as document layout analysis, visual information extraction, and document classification—while addressing challenges posed by diverse formats and layouts, ultimately supporting the digital transformation goals of organizations across multiple sectors [11, 33, 12].

6.4 Exploring Cross-domain and Real-world Applications

Exploring cross-domain and real-world applications of Document AI is essential for expanding its utility across various industries. Future research should focus on enhancing the DQN architecture and investigating task-specific features to improve document analysis performance, thus unlocking potential cross-domain applications [25]. By refining these architectures, Document AI can be tailored to meet the specific needs of domains such as legal, healthcare, and finance, where precise document analysis is critical.

Advancements in multi-modal models that integrate visual and textual information are crucial for processing entire documents in a single sequence, enhancing efficiency and accuracy in understanding [70]. This integration is particularly significant for complex documents requiring comprehensive understanding of both visual layouts and textual content.

To enhance Document AI applicability, expanding benchmarks to include advanced document understanding tasks and improving multimodal data integration are necessary steps [29]. This expansion allows for thorough evaluation of model capabilities and facilitates the development of robust solutions capable of handling diverse document types.

Additionally, integrating tools and techniques to improve the quality of generated rationales can extend Document AI applicability to multi-page document understanding [3]. This extension is vital for analyzing lengthy documents, such as contracts and reports, where understanding context across multiple pages is essential.

Research should also extend the examination of cognition and perception (CP) knowledge conflicts beyond document understanding to other multimodal tasks, such as scene understanding and visual reasoning [7]. Addressing these conflicts can enhance Document AI's interpretive capabilities and broaden its applicability.

Furthermore, exploring frameworks like VisToG in video-based scenarios could verify its effectiveness in processing dynamic and temporal data, broadening Document AI's scope to include video content analysis [41]. This exploration could open new avenues for Document AI in fields requiring video data analysis, such as surveillance and media.

By pursuing these research directions, Document AI can achieve greater cross-domain applicability, enabling accurate and efficient processing of documents in various real-world scenarios. This advancement is essential for improving Document AI technologies, which encompass automated reading, understanding, and analysis of diverse business documents. Leveraging deep learning methods for tasks such as document layout analysis and visual information extraction enhances

accuracy and efficiency, laying the groundwork for future innovations and broader applications across industries, ultimately supporting ongoing digital transformation and productivity increases [11, 12].

7 Conclusion

The exploration of multi-modal AI and document understanding technologies reveals their substantial role in enhancing the precision and efficiency of information extraction across various document formats. The synergy between large language models and visual elements has been instrumental in advancing form understanding tasks, as evidenced by the superior performance of transformer-based models that incorporate multi-modal inputs. Foundational datasets serve as critical resources, enabling future progress in this domain.

Innovative methodologies, such as the TRIE framework, demonstrate significant advantages, achieving superior efficiency and accuracy over existing state-of-the-art methods across multiple benchmarks. The strategic use of multiple pre-trained models for specific subtasks has further elevated the capabilities of multi-modal large language models, as shown by comprehensive experimental evaluations.

Despite these advancements, challenges remain, particularly in terms of dataset diversity, model calibration, and evaluation metrics, which are crucial for the effective deployment of multimodal document understanding models in practical applications. The integration of visual cues in Key Information Extraction systems and the exploration of generative methods to enhance extraction capabilities are identified as promising research directions.

The survey underscores the ongoing importance of addressing data quality and model adaptability to foster continued innovation in Document AI. Additionally, domain-specific frameworks, such as those leveraging synthetic annotations to integrate domain knowledge, offer a promising approach for achieving high performance with limited labeled data. These insights provide a foundation for future research aimed at overcoming existing limitations and expanding the applicability of document AI technologies across various sectors.

References

[1] Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. Multimodal pre-training based on graph attention network for document understanding, 2022.

[2] Haoqin Tu, Bingchen Zhao, Chen Wei, and Cihang Xie. Sight beyond text: Multi-modal training enhances llms in truthfulness and ethics, 2023.

[3] Wang Zhu, Alekh Agarwal, Mandar Joshi, Robin Jia, Jesse Thomason, and Kristina Toutanova. Efficient end-to-end visual document understanding with rationale distillation, 2024.

[4] Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. Omniparser: A unified framework for text spotting, key information extraction and table recognition, 2024.

[5] Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework, 2024.

[6] Tao Wu, Mengze Li, Jingyuan Chen, Wei Ji, Wang Lin, Jinyang Gao, Kun Kuang, Zhou Zhao, and Fei Wu. Semantic alignment for multimodal large language models, 2024.

[7] Zirui Shao, Chuwei Luo, Zhaoqing Zhu, Hangdi Xing, Zhi Yu, Qi Zheng, and Jiajun Bu. Is cognition consistent with perception? assessing and mitigating multimodal knowledge conflicts in document understanding, 2024.

[8] Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. Markuplm: Pre-training of text and markup language for visually-rich document understanding, 2022.

[9] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

[10] Rubèn Tito, Khanh Nguyen, Marlon Tobaben, Raouf Kerkouche, Mohamed Ali Souibgui, Kangsoo Jung, Joonas Jälkö, Vincent Poulain D'Andecy, Aurelie Joseph, Lei Kang, Ernest Valveny, Antti Honkela, Mario Fritz, and Dimosthenis Karatzas. Privacy-aware document visual question answering, 2024.

[11] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*, 2021.

[12] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document ai: Benchmarks, models and applications, 2021.

[13] Yihao Ding, Jean Lee, and Soyeon Caren Han. Deep learning based visually rich document content understanding: A survey, 2024.

[14] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*, 2024.

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

[16] Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms, 2023.

[17] Yongqiang Zhao, Zhenyu Li, Feng Zhang, Xinhai Xu, and Donghong Liu. Enhancing subtask performance of multi-modal large language model, 2023.

[18] Jian Lu, Shikhar Srivastava, Junyu Chen, Robik Shrestha, Manoj Acharya, Kushal Kafle, and Christopher Kanan. Revisiting multi-modal llm evaluation, 2024.

[19] Armineh Nourbakhsh, Sameena Shah, and Carolyn Rose. Towards a new research agenda for multimodal enterprise document understanding: What are we missing? *Findings of the Association for Computational Linguistics ACL 2024*, pages 14610–14622, 2024.

[20] Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanisławek. Document understanding dataset and evaluation (dude), 2023.

[21] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. A survey of deep learning approaches for ocr and document understanding, 2021.

[22] Zhengdong Lu, Xianggen Liu, Haotian Cui, Yukun Yan, and Daqi Zheng. Object-oriented neural programming (oonp) for document understanding, 2018.

[23] Abdelrahman Abdallah, Daniel Eberharter, Zoe Pfister, and Adam Jatowt. Transformers and language models in form understanding: A comprehensive review of scanned document analysis. *arXiv preprint arXiv:2403.04080*, 2024.

[24] Zhen Qin, Daoyuan Chen, Wenhao Zhang, Liuyi Yao, Yilun Huang, Bolin Ding, Yaliang Li, and Shuiguang Deng. The synergy between data and multi-modal large language models: A survey from co-development perspective, 2024.

[25] Sumit Shekhar, Bhanu Prakash Reddy Guda, Ashutosh Chaubey, Ishan Jindal, and Avneet Jain. Opad: An optimized policy-based active learning framework for document content analysis, 2021.

[26] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing, 2023.

[27] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. Docformerv2: Local features for document understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 709–718, 2024.

[28] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding, 2024.

[29] Liang Qiao, Hui Jiang, Ying Chen, Can Li, Pengfei Li, Zaisheng Li, Baorui Zou, Dashan Guo, Yingda Xu, Yunlu Xu, Zhanzhan Cheng, and Yi Niu. Davarocr: A toolbox for ocr and multi-modal document understanding, 2022.

[30] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer, 2022.

[31] Pengfei Hu, Zhenrong Zhang, Jiefeng Ma, Shuhang Liu, Jun Du, and Jianshu Zhang. Docmamba: Efficient document pre-training with state space model, 2025.

[32] Allison Hegel, Marina Shah, Genevieve Peaslee, Brendan Roof, and Emad Elwany. The law of large documents: Understanding the structure of legal contracts using visual cues, 2021.

[33] Sotirios Kastanas, Shaomu Tan, and Yi He. Document ai: A comparative study of transformer-based, graph-based models, and convolutional neural networks for document layout analysis, 2023.

[34] Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. Trie: End-to-end text reading and information extraction for document understanding, 2021.

[35] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 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*, volume 36, pages 10767–10775, 2022.

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

[37] Marcel Lamott and Muhammad Armaghan Shakir. Leveraging distillation techniques for document understanding: A case study with flan-t5, 2024.

[38] Anjanava Biswas and Wrick Talukdar. Robustness of structured data extraction from in-plane rotated documents using multi-modal large language models (llm), 2024.

[39] Kilian Carolan, Laura Fennelly, and Alan F. Smeaton. A review of multi-modal large language and vision models, 2024.

[40] Yihao Ding, Lorenzo Vaiani, Caren Han, Jean Lee, Paolo Garza, Josiah Poon, and Luca Cagliero. 3mvrd: Multimodal multi-task multi-teacher visually-rich form document understanding, 2024.

[41] Minbin Huang, Runhui Huang, Han Shi, Yimeng Chen, Chuanyang Zheng, Xiangguo Sun, Xin Jiang, Zhenguo Li, and Hong Cheng. Efficient multi-modal large language models via visual token grouping, 2024.

[42] nd-to-end text reading and infor.

[43] Jordy Van Landeghem, Subhajit Maity, Ayan Banerjee, Matthew Blaschko, Marie-Francine Moens, Josep Lladós, and Sanket Biswas. Distildoc: Knowledge distillation for visually-rich document applications, 2024.

[44] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding, 2024.

[45] Souhail Bakkali, Sanket Biswas, Zuheng Ming, Mickaël Coustaty, Marçal Rusiñol, Oriol Ramos Terrades, and Josep Lladós. Globaldoc: A cross-modal vision-language framework for real-world document image retrieval and classification. *arXiv preprint arXiv:2309.05756*, 2023.

[46] Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18733–18741, 2024.

[47] Hong Chen, Xin Wang, Yuwei Zhou, Bin Huang, Yipeng Zhang, Wei Feng, Houlun Chen, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Multi-modal generative ai: Multi-modal llm, diffusion and beyond. *arXiv preprint arXiv:2409.14993*, 2024.

[48] Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. Lvlm-interpret: An interpretability tool for large vision-language models, 2024.

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

[50] Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2d documents, 2018.

[51] Xin He, Longhui Wei, Lingxi Xie, and Qi Tian. Incorporating visual experts to resolve the information loss in multimodal large language models. *arXiv preprint arXiv:2401.03105*, 2024.

[52] Haoyu Cao, Changcun Bao, Chaohu Liu, Huang Chen, Kun Yin, Hao Liu, Yinsong Liu, Deqiang Jiang, and Xing Sun. Attention where it matters: Rethinking visual document understanding with selective region concentration, 2023.

[53] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022.

[54] Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. Hrvda: High-resolution visual document assistant, 2024.

[55] Haoli Bai, Zhiguang Liu, Xiaojun Meng, Wentao Li, Shuang Liu, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, Xin Jiang, and Qun Liu. Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding, 2022.

[56] Jiapeng Wang, Lianwen Jin, and Kai Ding. Lilt: A simple yet effective language-independent layout transformer for structured document understanding, 2022.

[57] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Structurallm: Structural pre-training for form understanding. *arXiv preprint arXiv:2105.11210*, 2021.

[58] Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. End-to-end document recognition and understanding with dessurt, 2022.

[59] Anna Scius-Bertrand, Michael Jungo, Lars Vögtlin, Jean-Marc Spat, and Andreas Fischer. Zero-shot prompting and few-shot fine-tuning: Revisiting document image classification using large language models, 2024.

[60] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning, 2023.

[61] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645, 2023.

[62] Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. Visually-situated natural language understanding with contrastive reading model and frozen large language models, 2023.

[63] Tianlong Ma, Xingjiao Wu, Xin Li, Xiangcheng Du, Zhao Zhou, Liang Xue, and Cheng Jin. Document layout analysis with aesthetic-guided image augmentation, 2021.

[64] Zhiming Mao, Haoli Bai, Lu Hou, Jiansheng Wei, Xin Jiang, Qun Liu, and Kam-Fai Wong. Visually guided generative text-layout pre-training for document intelligence. *arXiv preprint arXiv:2403.16516*, 2024.

[65] Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. Docvlm: Make your vlm an efficient reader, 2024.

[66] Lijun Yu, Jin Miao, Xiaoyu Sun, Jiayi Chen, Alexander G. Hauptmann, Hanjun Dai, and Wei Wei. Documentnet: Bridging the data gap in document pre-training, 2023.

[67] Jun Chen, Dannong Xu, Junjie Fei, Chun-Mei Feng, and Mohamed Elhoseiny. Document haystacks: Vision-language reasoning over piles of 1000+ documents. *arXiv preprint arXiv:2411.16740*, 2024.

[68] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*, 2024.

[69] Loris Giulivi and Giacomo Boracchi. Explaining multi-modal large language models by analyzing their vision perception, 2024.

[70] Thibault Douzon, Stefan Duffner, Christophe Garcia, and Jérémy Espinas. Long-range transformer architectures for document understanding, 2023.

[71] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R. Manmatha. Docformerv2: Local features for document understanding, 2023.

[72] Hsiu-Wei Yang, Abhinav Agrawal, Pavlos Fragkogiannis, and Shubham Nitin Mulay. Can ai models appreciate document aesthetics? an exploration of legibility and layout quality in relation to prediction confidence, 2024.

[73] Thomas Delteil, Edouard Belval, Lei Chen, Luis Goncalves, and Vijay Mahadevan. Matrix – modality-aware transformer for information extraction, 2022.

[74] Shrey Mishra, Antoine Gauquier, and Pierre Senellart. Modular multimodal machine learning for extraction of theorems and proofs in long scientific documents (extended version), 2024.

[75] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.

[76] I. de Rodrigo, A. Sanchez-Cuadrado, J. Boal, and A. J. Lopez-Lopez. The merit dataset: Modelling and efficiently rendering interpretable transcripts, 2024.

[77] Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. Llasm: Large language and speech model, 2023.

[78] Łukasz Borchmann. Notes on applicability of gpt-4 to document understanding, 2024.

[79] Ran Zmigrod, Pranav Shetty, Mathieu Sibue, Zhiqiang Ma, Armineh Nourbakhsh, Xiaomo Liu, and Manuela Veloso. "what is the value of templates?" rethinking document information extraction datasets for llms, 2024.

[80] Lee Kezar and Jay Pujara. Finding pragmatic differences between disciplines, 2023.

[81] Chuanghao Ding, Xuejing Liu, Wei Tang, Juan Li, Xiaoliang Wang, Rui Zhao, Cam-Tu Nguyen, and Fei Tan. Synthdoc: Bilingual documents synthesis for visual document understanding, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.