Multimodal Large Language Models for Text-rich Image Understanding: A Comprehensive Review

Pei Fu[1], Tongkun Guan[2], Zining Wang[1], Zhentao Guo[3], Chen Duan[1], Hao Sun[4], Boming Chen[1], Jiayao Ma[1], Qianyi Jiang[1], Kai Zhou[1], Junfeng Luo[1]

[1]Meituan, [2] Shanghai Jiao Tong University, [3]Beijing Institute of Technology, [4]MAIS & NLPR, Institute of Automation, Chinese Academy of Sciences

{fupei,duanchen02,wangzining03,chenboming,majiayao02}@meituan.com
{jiangqianyi02,zhoukai03,luojunfeng}@meituan.com
gtk0615@sjtu.edu.cn,hao.sun@cripac.ia.ac.cn,zt_guo1230@163.com

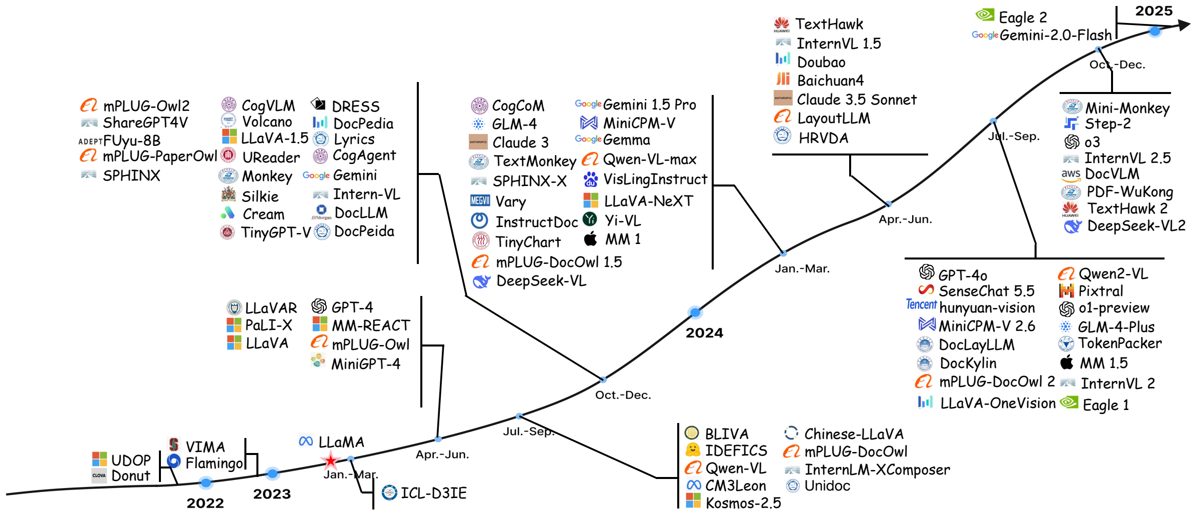

Figure 1: The development timeline of TIU MLLMs.

Abstract

The recent emergence of Multi-modal Large Language Models (MLLMs) has introduced a new dimension to the Text-rich Image Understanding (TIU) field, with models demonstrating impressive and inspiring performance. However, their rapid evolution and widespread adoption have made it increasingly challenging to keep up with the latest advancements. To address this, we present a systematic and comprehensive survey to facilitate further research on TIU MLLMs. Initially, we outline the timeline, architecture, and pipeline of nearly all TIU MLLMs. Then, we review the performance of selected models on mainstream benchmarks. Finally, we explore promising directions, challenges, and limitations within the field.

1 Introduction

Text-rich images play a pivotal role in real-world scenarios by efficiently conveying complex information and improving accessibility (Biten et al., 2019). Accurately interpreting these images is essential for automating information extraction, advancing AI systems, and optimizing user interactions. To formalize this research domain, we term it **T**ext-rich **I**mage **U**nderstanding (**TIU**), which encompasses two core capabilities: perception and understanding. The perception dimension focuses on visual recognition tasks, such as text detection (Liao et al., 2022), text recognition (Guan et al., 2025), formula recognition (Truong et al., 2024; Guan et al., 2024a), and document layout analysis (Yupan et al., 2022). The understanding dimension, conversely, requires semantic reasoning for applications like key information extraction and document-based visual question answering (*e.g.*, DocVQA (Mathew et al., 2021b), ChartQA (Masry et al., 2022), and TextVQA (Singh et al., 2019)).

Historically, perception and understanding tasks were handled separately through specialized models or multi-stage pipelines. Recent advances in vision-language models have unified these tasks within Visual Question Answering (VQA) paradigms, driving research towards the development of end-to-end universal models.

Figure 1 presents an evolutionary timeline delineating critical milestones in unified text-rich image understanding models. The trajectory reveals two distinct eras: (a) The pre-LLM period (2019-

Findings of the Association for Computational Linguistics: ACL 2025, pages 19941–19958
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

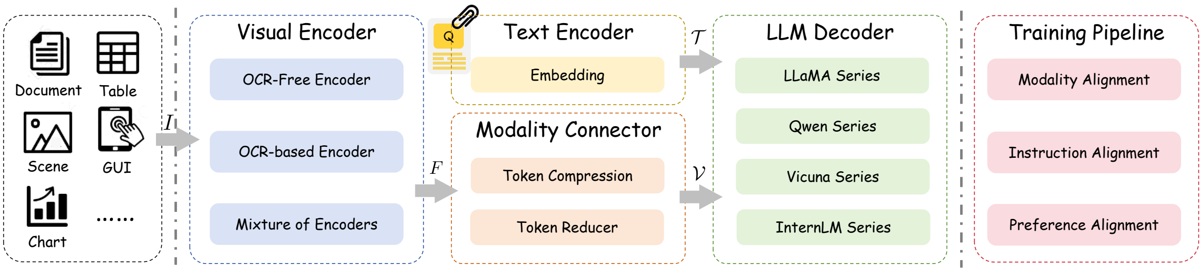

Figure 2: The general model architecture of MLLMs and the implementation choices for each component.

2022) characterized by specialized architectures like LayoutLM (Xu et al., 2019) and Donut (Kim et al., 2021), which employed modality-specific pre-training objectives (masked language modeling, masked image modeling, *etc.*) coupled with OCR-derived supervision (text recognition, spatial order recovery, *etc.*). While effective in controlled settings, these models exhibited limited adaptability to open-domain scenarios due to their task-specific fine-tuning requirements and constrained cross-modal interaction mechanisms. (b) The post-LLM era (2023–present) is marked by the growing popularity of LLMs. Some studies propose Multimodal Large Language Models (MLLMs), which integrate LLM with visual encoders to jointly process visual tokens and linguistic elements through unified attention mechanisms, achieving end-to-end sequence modeling.

This paradigm evolution addresses two critical limitations of earlier methods. First, the emergent MLLM framework eliminates modality-specific inductive biases through homogeneous token representation, enabling seamless multi-task integration. Second, the linguistic priors encoded in LLMs empower unprecedented zero-shot generalization and allow direct application to diverse tasks without task-specific tuning.

Although these MLLMs present impressive and inspiring results, their rapid evolution and broad adoption have made tracking cutting-edge advancements increasingly challenging. Therefore, a systematic review that is tailored for documents to summarize and analyse these methods is in demand. However, existing surveys on text-rich image understanding often exhibit narrow focus: they either analyze domain-specific scenarios (e.g., tables and figures (Huang et al., 2024a), charts (Huang et al., 2024b; Al-Shetairy et al., 2024), forms (Abdallah et al., 2024)) or emphasize unified deep learning frameworks (Subramani et al.; Ding et al., 2024).

Our systematic survey addresses the gap by providing the first comprehensive analysis of nearly all TIU MLLMs in four dimensions: Model Architectures (Section 2), Training Pipeline (Section 3), Datasets and Benchmarks (Section 4), Challenges and Trends (Section 5). This holds both academic and practical significance for advancing the field (Ma et al., 2024a).

2 Model Architecture

TIU MLLM methods typically leverage pre-trained general visual foundation models to extract robust visual features or employ OCR engines to capture text and layout information from images. A modality connector is then used to align these visual features with the semantic space of the language features from the LLM. Finally, the combined visual-language features are fed into the LLM, which utilizes its powerful comprehension capabilities for semantic reasoning to generate the final answer. As illustrated in Figure 2, the framework of TIU MLLMs can be abstracted into three core components: Visual Encoder, Modality Connector, and LLM Decoder.

2.1 Visual Encoder

The Visual Encoder $\mathcal{F}(\cdot)$ is responsible for transforming input image $\mathbf{I}$ into feature representations $\mathbf{V}$, expressed as $\mathbf{V} = \mathcal{F}(\cdot)(\mathbf{I})$. As illustrated in Figure 3, these encoders can be categorized into OCR-free, OCR-based, or a hybrid approach.

OCR-free Encoder is widely used to extract high-level visual features, effectively capturing essential information about objects, scenes, and textures. The commonly used OCR-free encoders include (1) **CLIP** (Radford et al., 2021); (2) **ConvNeXt** (Woo et al., 2023); (3) **SAM** (Kirillov et al., 2023); (4) **DINOv2** (Oquab et al., 2023); (5) **Swin-T** (Liu et al., 2021); (6) **InternViT** (Chen et al., 2024d).

OCR-based Encoder processes textual content and layout information from OCR outputs through three primary paradigms: (1) **Direct Input** injects raw OCR texts into LLMs, though long sequences degrade inference efficiency (He et al., 2023b);

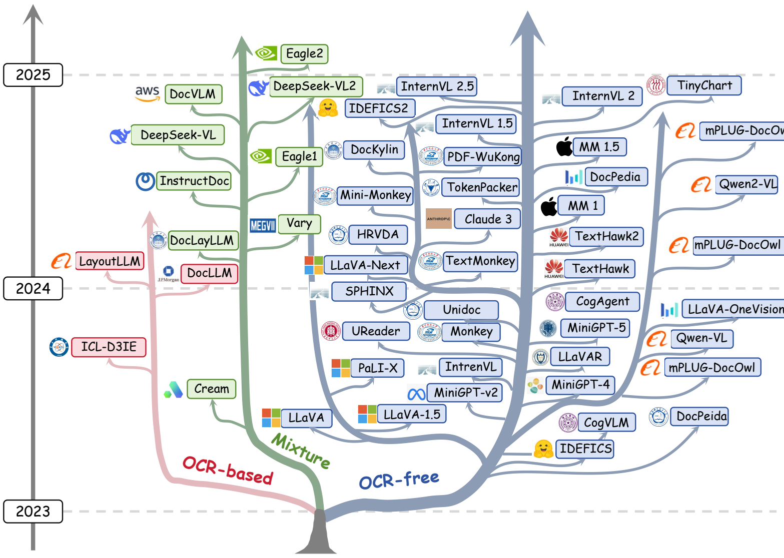

Figure 3: The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. According to the classification of Encoders, the blue branch is ocr-free, the pink branch is ocr-based, and the green branch is Mixture of Encoders.

(2) **Cross-Attention** dynamically selects salient content via attention mechanisms within LLMs (Wang et al., 2023); (3) **External Encoder** employs specialized models like BLIP-2 (Li et al., 2023), DocFormerv2 (Nacson et al., 2024) or LayoutLMv3 (Yupan et al., 2022) to structure OCR features before LLM integration (Tanaka et al., 2024a; Luo et al., 2024a; Fujitake, 2024).

Mixture of Encoders strategies address TIU task complexity through two dominant configurations: (1) **Dual OCR-Free** architectures (*e.g.*, CLIP+SAM) combine complementary visual encoders to jointly capture global semantics and local details (Wei et al., 2024); (2) **Hybrid OCR-Free/OCR-Based** systems (*e.g.*, CLIP+LayoutLMv3) synergize visual feature extraction with text-layout understanding, proving particularly effective for document-level tasks requiring multimodal reasoning (Liao et al., 2024a).

2.2 Modality Connector

Visual embeddings $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n]$ and language embeddings $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_l]$ belong to different modalities. Consequently, to bridge the gap between them and create unified sequence representations that can be processed by large language models (LLMs), a modality connector $\xi : \mathbf{V} \rightarrow \mathbf{T}$ is typically employed, which is responsible for converting n visual features into m visual tokens. We review the strategies previously utilized in the literature for this purpose.

Specifically, the modality connector can be easily implemented using a simple linear projector or multi-layer perception (MLP), *i.e.*, $m = n$, but faces challenges in scalability and efficiency. Recent works also proposed more effective and innovative modality connectors from various perspectives, such as token compression and token reduction. The former focuses on reducing the number of inputs to the MLLM token with lossless compression, and the latter addresses the issue of costly tokens by removing redundant and unimportant token representations, such as background tokens.

Token Compression

1) Pixel shuffle (Chen et al., 2024c) rearranges the elements of a high-resolution feature map (h, w) to form a lower-resolution feature map $(\frac{h}{s}, \frac{w}{s})$ by redistributing the spatial dimensions into the depth (channels) of the feature map. Here, s denotes the compression rate. We summarized the process as $\xi : \mathbb{R}^{h\times w\times C} \rightarrow \mathbb{R}^{\frac{h}{s}\times\frac{w}{s}\times(s\times C)}$.

Token Reducer

1) Cross Attention (Alayrac et al., 2022; Li et al., 2023; Chen et al., 2024d; Dai et al., 2023) operates on the queries (a group of trainable vectors or the key features of the model itself) and the keys which are the image features produced by the vision encoder. We summarized the process as $\xi : \mathbb{R}^{h\times w\times C} \rightarrow \mathbb{R}^{q\times D}$.

2) H-Reducer (Hu et al., 2024a) introduces the 1×4 convolution layer to reduce visual features, as the horizontal texts are widely found in natural scenes and semantically coherent. We summarized the process as $\xi : \mathbb{R}^{h\times w\times C} \rightarrow \mathbb{R}^{h\times\frac{w}{4}\times D}$.

3) C/D-abstract (Cha et al., 2024) employs Convolution and Deformable Attention respectively to

Model	Visual Encoder	Modality Connector	LLM Decoder	Training Pipline	DocVQA	InfoVQA	ChartQA	TextVQA	Avg.
UReader (Ye et al., 2023b)	CLIP-ViT-L/14	Cross Attention	LLaMA-7B	MA+IA	65.4	42.2	59.3	57.6	56.13
DocLLM-1B (Wang et al., 2023)	-	-	Falcon-1B	MA+IA	61.4	-	-	-	-
DocLLM-7B (Wang et al., 2023)	-	-	LLaMA2-7B	MA+IA	69.5	-	-	-	-
Cream (Kim et al., 2023)	CLIP-ViT-L/14	Cross Attention	Vicuna-7B	MA+IA	79.5	43.5	63.0	-	-
LLaVA-13B (Liu et al., 2023c)	CLIP-ViT-L/14	MLP	Vicuna-13B	MA+IA	6.9	-	-	36.7	-
PaLI-X (Chen et al., 2023)	ViT-22B	MLP	UL2-32B	MA+IA	86.8	54.8	72.3	80.8	73.68
LLaVAR (Zhang et al., 2023)	CLIP-ViT-L/14	MLP	Vicuna-13B	MA+IA	11.6	-	-	48.5	-
Qwen-VL (Bai et al., 2023b)	ViT-bigG	Cross Attention	Qwen-7B	MA+IA	65.1	35.4	65.7	63.8	57.50
LLaVA-1.5-7B (Liu et al., 2023b)	CLIP-ViT-L	MLP	Vicuna1.5-7B	MA+IA	-	-	-	58.2	-
LLaVA-1.5-13B (Liu et al., 2023b)	CLIP-ViT-L	MLP	Vicuna1.5-13B	MA+IA	-	-	-	62.5	-
CogAgent (Hong et al., 2023)	EVA2-CLIP	MLP+Cross Attention	Vicuna-13B	MA+IA	81.6	44.5	68.4	76.1	67.65
Unidoc (Feng et al., 2023)	CLIP-ViT-L/14	MLP	Vicuna-13B	MA+IA	90.2	36.8	70.5	73.7	67.80
Monkey (Li et al., 2024e)	Vit-BigG	Cross Attention	Qwen-7B	MA+IA	66.5	36.1	65.1	67.6	58.83
Mini-Monkey (Huang et al., 2024c)	InternViT-300M	MLP	InternLLM2-2B	IA	87.4	60.1	76.5	75.7	74.93
TextMonkey (Liu et al., 2024e)	Vit-BigG	Cross Attention	Qwen-7B	MA+IA	73.0	-	66.9	65.6	-
IDEFICS2 ((Laurençon et al., 2024))	SigLIP-SO400M	Cross Attention	Mistral-7B	MA+IA	74.0	-	-	73.0	-
LayoutLLM (Luo et al., 2024b)	LayoutLMv3-large	MLP	Vicuna1.5-7B	MA+IA	74.25	-	-	-	-
DocKylin (Zhang et al., 2024b)	Swin	MLP	Qwen-7B	MA+IA	77.3	46.6	66.8	-	-
DocLayLLM (Liao et al., 2024b)	LayoutLMV3	MLP	LLaMA3-8B	MA+IA	77.79	42.02	-	-	-
mPLUG-DocOwl (Hu et al., 2024a)	CLIP-ViT-L/14	Cross Attention	LLaMA-7B	MA+IA	62.2	38.2	57.4	52.6	52.60
mPLUG-DocOwl1.5 (Hu et al., 2024b)	CLIP-ViT-L/14	H-Reducer	LLaMA2-7B	MA+IA	82.2	50.7	70.2	68.6	67.93
mPLUG-DocOwl2 (Hu et al., 2024d)	CLIP-ViT-L/14	H-Reducer	LLaMA2-7B	MA+IA	80.7	46.4	70.0	66.7	65.95
Vary (Wei et al., 2024)	CLIP-ViT-L/14 + SAM	MLP	Qwen-7B	MA+IA	76.3	-	66.1	-	-
Eagle (Shi et al., 2024)	CLIP + ConvNeXt + Pix2Struct + EVA2 + SAM	MLP	LLaMA3-8B	MA+IA	86.6	-	80.1	77.1	-
PDF-WuKong (Xie et al., 2024)	CLIP-ViT-L-14	Cross Attention	InernLM2-7B	MA+IA	85.1	61.3	80.0	-	-
InstructDoc (Tanaka et al., 2024b)	CLIP/Eva-CLIP-ViT	Cross Attention + MLP	Flan-T5/OPT	MA+IA	-	50.9	29.4	53.8	-
TextHawk (Yu et al., 2024a)	SigLIP	Cross Attention	InternLM-XComposer	MA+IA	76.4	50.6	66.6	-	-
TextHawk2 (Yu et al., 2024b)	SigLIP	Cross Attention	Qwen2-7B	MA+IA	89.6	67.8	81.4	75.1	78.48
MM1.5-1B (Zhang et al., 2024a)	CLIP-ViT-H	C-Abstractor	Private	MA+IA	81.0	50.5	67.2	72.5	67.80
MM1.5-3B (Zhang et al., 2024a)	CLIP-ViT-H	C-Abstractor	Private	MA+IA	87.7	58.5	74.2	76.5	74.23
MM1.5-7B (Zhang et al., 2024a)	CLIP-ViT-H	C-Abstractor	Private	MA+IA	88.1	59.5	78.6	76.5	75.68
MM1.5-30B (Zhang et al., 2024a)	CLIP-ViT-H	C-Abstractor	Private	MA+IA	91.4	67.3	83.6	79.2	80.38
HRVDA (Liu et al., 2024a)	Swin-L	MLP	LLaMA2-7B	MA+IA	72.1	43.5	67.6	73.3	64.13
InternVL1.5-26B (Chen et al., 2024c)	InternViT-6B	Pixel-shuffle + MLP	InternLM2-20B	MA+IA	90.9	72.5	83.8	80.6	81.95
InternVL2.5-1B (Chen et al., 2024b)	InternViT-300M	Pixel-shuffle + MLP	Qwen2.5-0.5B	MA+IA	84.8	56.0	75.9	72.0	72.18
InternVL2.5-2B (Chen et al., 2024b)	InternViT-300M	Pixel-shuffle + MLP	InternLM2.5-1.8B	MA+IA	88.7	60.9	79.2	74.3	75.78
InternVL2.5-4B (Chen et al., 2024b)	InternViT-300M	Pixel-shuffle + MLP	Qwen2.5-3B	MA+IA	91.6	72.1	84.0	76.8	81.13
InternVL2.5-8B (Chen et al., 2024b)	InternViT-300M	Pixel-shuffle + MLP	InternLM2.5-7B	MA+IA	93.0	77.6	84.8	79.1	83.63
InternVL2.5-26B (Chen et al., 2024b)	InternViT-6B	Pixel-shuffle + MLP	InternLM2.5-20B	MA+IA	94.0	79.8	87.2	82.4	85.85
InternVL2.5-38B (Chen et al., 2024b)	InternViT-6B	Pixel-shuffle + MLP	Qwen2.5-32B	MA+IA	95.3	83.6	88.2	82.7	87.45
InternVL2.5-78B (Chen et al., 2024b)	InternViT-6B	Pixel-shuffle + MLP	Qwen2.5-72B	MA+IA	95.1	84.1	**88.3**	83.4	87.73
InternVL2.5-8B-mpo(Wang et al., 2024c)†	InternViT-300M	Pixel-shuffle +MLP	InternLM2.5-7B	PA	92.3	76.0	83.8	79.1	82.80
DocPedia (Feng et al., 2024)	Swin	MLP	Vicuna-7B	MA+IA	47.1	15.2	46.9	60.2	42.35
TinyChart (Zhang et al., 2024d)	SigLIP	MLP	Phi-2	IA	-	-	83.6	-	-
TokenPacker-7B (Li et al., 2024d)	CLIP-ViT-L/14	Cross Attention	Vicuna-7B	MA+IA	60.2	-	-	-	-
TokenPacker-13B (Li et al., 2024d)	CLIP-ViT-G/14	Cross Attention	Vicuna-13B	MA+IA	70.0	-	-	-	-
LLaVA-OneVision-0.5B (Li et al., 2024a)	SigLIP	MLP	qwen2-0.5B	MA+IA	70.0	41.8	61.4	-	-
LLaVA-OneVision-7B (Li et al., 2024a)	SigLIP	MLP	qwen2-7B	MA+IA	87.5	68.8	80.0	-	-
Qwen2-VL-2B (Wang et al., 2024b)	CLIP-VIT-G/14	Cross Attention	Qwen2-2B	MA+IA	90.1	65.5	73.5	79.7	77.20
Qwen2-VL-7B (Wang et al., 2024b)	CLIP-VIT-G/14	Cross Attention	Qwen2-7B	MA+IA	94.5	76.5	83.0	84.3	84.58
DocVLM (Nacson et al., 2024)	CLIP-ViT-G/14 + DocFormerV2	Cross Attention	Qwen2-7B	MA+IA	92.8	66.8	-	82.8	-
Qwen2-VL-72B (Wang et al., 2024b)	CLIP-VIT-G/14	Cross Attention	Qwen2-72B	MA+IA	**96.5**	**84.5**	**88.3**	**85.5**	**88.70**
DeepSeek-VL2-3B (Wu et al., 2024b)	SigLIP-SO400M-384	Pixel-shuffle + MLP	DeepSeekMoE	MA+IA	88.9	66.1	81.0	80.7	79.18
DeepSeek-VL2-16B (Wu et al., 2024b)	SigLIP-SO400M-384	Pixel-shuffle + MLP	DeepSeekMoE	MA+IA	92.3	75.8	84.5	83.4	84.00
DeepSeek-VL2-27B (Wu et al., 2024b)	SigLIP-SO400M-384	Pixel-shuffle + MLP	DeepSeekMoE	MA+IA	93.3	78.1	86.0	84.2	85.40
Eagle2 (Li et al., 2025b)	SigLIP + ConvNeXt	MLP	Qwen2.5-7B	MA+IA	92.6	77.2	86.4	83.0	84.80

Table 1: The summary of representative mainstream MLLMs, including the model architectures, training pipelines, and scores on the four most popular benchmarks of TIU. "Private" indicates that the MLLM utilizes a proprietary large model. "†" indicates the results are obtained by downloading official open-source model and testing it locally.

achieve both flexibility and locality preservation.

4) Attention Pooling (Liu et al., 2024e; Huang et al., 2024c) identifies important tokens and removes redundant ones. To evaluate the redundancy of image features, the similarity between image tokens is often utilized (Liu et al., 2024e). This method selects tokens that are highly unique and lack closely similar counterparts. Average pooling is the most special one.

2.3 LLM Decoder

The aligned features are fed into the LLM decoder together with the language embeddings for reasoning. We list the commonly used LLMs in MLLM:

LLaMA Series. LLaMA (Touvron et al., 2023a,b; Dubey et al., 2024) is a series of open-source large language models developed by Meta, aimed at promoting openness and innovation in artificial intelligence technology, LLaMA series include models of varying parameter scales (e.g., 7B, 13B, 34B).

Qwen Series. Qwen (Bai et al., 2023a; Yang et al., 2024a), developed by Alibaba, is a multilingual LLM that supports both Chinese and English.

Vicuna Series. Vicuna (Zheng et al., 2023) is an open-source large language model built on LLaMA, developed by research teams from institutions including UC Berkeley, CMU, and Stanford.

InternLM series. InternLM (Cai et al., 2024) is an open-source large language model series developed by the Shanghai Artificial Intelligence Laboratory, with the latest version, InternLM 2.5, offering parameter sizes of 1.8B, 7B, and 20B.

The role of the LLM decoder in MLLM for rich text image understanding can be summarized in the following four points: 1) Core Functionality: The LLM serves as the reasoning engine in MLLMs, integrating visual features (from encoders) and users' instructions to generate coherent responses. 2) Tokenizer: Early tokenizers often faced limitations with rare or multilingual text in images. Modern MLLMs address this by leveraging expansive subword vocabularies, covering diverse languages, thereby improving robustness for text-rich scenarios. 3) Multilingual Capabilities: Pretrained multilingual LLM directly enhance MLLMs'cross-lingual text comprehension in images. 4) Context Length: Long-context decoders enable MLLMs to process dense textual layouts or text-rich images by maintaining spatial and semantic relationships across extended sequences.

3 Training Pipeline

The training pipeline of MLLM for TIU can be delineated into three main stages: 1) Modality Alignment (**MA**); 2) Instruction Alignment (**IA**); and 3) Preference Alignment (**PA**).

3.1 Modality Alignment

In this stage, previous works typically use OCR data from traditional OCR tasks (Guan et al., 2024b, 2023b) to pre-train the MLLM, which aims to bridge the modality gap. The general alignment methods can be categorized into three types: recognition, localization, and parsing.

Read Full Text. UReader (Ye et al., 2023b) is the first to explore unified document-level understanding, which introduces the Read Full Text task in VQA for pre-training. Specifically, they include 1) reading all texts from top to bottom and left to right, and 2) reading the remaining texts based on given texts. Compared to reading the full text, some works (Lv et al., 2024; Hu et al., 2024a) proposes a more structured reading approach by predicting the image markdown, not text transcriptions.

Reading Partial Text within Localization. Due to the length of document texts, instructions for reading the full text may risk truncation because of the limited token length in LLMs. To address these limitations, Park *et al.*(Park et al., 2024) introduced two novel tasks: Reading Partial Text (RPT) and Predicting Text Position (PTP). The former randomly selects and reads continuous portions of text in the reading order from top to bottom and left to right. For example, "Q: What is the text in the image between the first 30%, from 20% to 40%, or the last 16%?" For the PTP task, given a text segment, the MLLM aims to infer its relative position (percentage format) within the full text. For example, "Q: Where is the text query texts located within the image? A: 40% to 80%". However, this approach can be somewhat obscure and challenging to express accurately.

Alternatively, some methods (Hu et al., 2024b; Yu et al., 2024a; Liu et al., 2024a) extract texts based on specific spatial positions, which are summarized into two types. 1) Text Recognition aims to extract the textual content from a given position in the image, ensuring that the model can accurately recognize and extract text within specific regions. 2) Text Grounding involves identifying the corresponding bounding box for specific text in the image, which assists the model in understanding the document layout.

Parsing. In document images, many elements (charts, formulas, and tables) may not be represented using plain text. An increasing number of researchers are now focusing on these element parsing. 1) Chart Parsing. Chart types include vertical bars, horizontal bars, lines, scatter plots, and pie charts. Charts serve as visual representations of tables, and organizing text in reading order fails to capture their structure. To preserve their mathematical properties, researchers often convert charts into tables. This process involves breaking down the chart into x/y axes and their corresponding values, which can be represented in Markdown, CSV formats, or even converted into Python code. This approach enables models to better understand the chart's specific meaning.

2) Table Parsing. Compared to charts, tables have a more standardized structure, where rows and columns form key-value pairs. Common formats for representing tables include LaTeX, Markdown, and HTML. Markdown is often used for simple tables due to its concise text format, while HTML can handle cells that span multiple rows and columns, despite its use of many paired tags like <tr></tr> and <td></td>. Some tables, with complex spanning, custom lines, spacing, or multi-page length, require LaTeX for representation. However, the diversity in LaTeX representations can make these tables challenging for models to fully understand.

3) Formula Parsing. Besides tables and charts, formulas are also commonly used. In the pretraining phase, models learn the LaTeX representation of formula images, enhancing their understanding of formulas. This provides a solid foundation for tasks involving formula computation and reasoning during the instruction alignment.

3.2 Instruction Alignment

Upon completing the modality alignment pre-training stage, the MLLM acquires basic visual recognition and dialogue capabilities. However, to achieve human-aligned intelligence, three critical capability gaps must be addressed:(1) Advanced multimodal perception and cross-modal reasoning abilities; (2) Prompt robustness across diverse formulations; (3) Zero-shot generalization for unseen task scenarios. To bridge these gaps, instruction alignment through supervised fine-tuning (SFT) has emerged as an effective paradigm. This phase typically unfreezes all model parameters and employs instructional data with structured templates.

To systematically address these challenges, we have categorized the current methods emerging in instruction alignment into three distinct levels:

1) Level 1: Visual-Semantic Anchoring. We categorize these instructions into two types: i) Answer within the image; and ii) Answer without the image. This type of instruction data where answers are located directly within the image, assists MLLMs refine their accuracy in generating responses that are directly linked to specific visual content, reducing reliance on generic or contextually weak answers (Mathew et al., 2021b, 2022). Certain tasks require reasoning based on world knowledge and involve complex inference procedures, such as scientific question answering (Masry et al., 2022; Chen et al., 2021). Consequently, these instructions are designed with the common characteristic that the answer is not directly visible in the image. This encourages the model to utilize its linguistic comprehension and external knowledge, enhancing its advanced reasoning and inference capabilities. An example might be: "Q: How much higher is the red bar compared to the yellow bar in the chart, in terms of percentage? A: 12.1%."

2) Level 2: Prompt Diversity Augmentation. To bolster robustness in handling a broader spectrum of prompts, rather than being limited to specific prompts tailored for particular tasks, researchers often employ data augmentation on the question component of the instruction stream. A popular strategy involves leveraging existing large language models to rephrase the same question in multiple ways. For example, consider the original question: "What is written on the sign in the image?" It can be rephrased as: "Can you read the text displayed on the sign shown in the image?""Identify the sign in the image.""Please examine the image and list the words that appear within the sign." By utilizing such varied templates, researchers can train MLLMs to better interpret and respond to a wide range of prompts, thereby enhancing their flexibility and accuracy in real-world applications.

3) Level 3: Zero-shot Generalization. To enhance the generalization ability to handle unseen tasks, several strategies typically are employed:

Chain of Thought (CoT) (Wei et al., 2022) reasoning involves breaking down complex problems into a series of intermediate steps or sub-tasks, allowing a model to tackle each part systematically. Some studies have demonstrated improvements by incorporating text-level CoT reasoning (Zhang et al., 2024c) or box-level visual CoT supervision (Shao et al., 2025). To better illustrate the process, consider the prompt: "What is the average of the last four countries' data?", the CoT reasoning unfolds as follows: i) Identify the data for the last four countries; ii) Calculate the sum of these values; iii) Calculate the average by dividing the sum by the number of countries.

Another strategy is Retrieval-Augmented Generation (RAG). RAG (Arslan et al., 2024) combines the strengths of retrieval-based and generation-based approaches by integrating an information retrieval component with a generative model. This method allows the model to access a vast external knowledge base, retrieving pertinent information to inform and enhance the generation process.

3.3 Preference Alignment

In the modality and instruction alignment stages, the model predicts the next token based on previous ground-truth tokens during training, and on its own prior outputs during inference. If errors occur in the outputs, this can lead to a distribution shift in inference. The more output the model has, the more serious this phenomenon becomes. In previous natural language processing (NLP) works (Lai et al., 2024; Pang et al., 2025), a series of preference alignment techniques (Rafailov et al., 2024; Ouyang et al., 2022; Shao et al., 2024; Wang et al., 2024a) have been proposed to optimize the output of the model to make it more consistent with human values and expectations. Benefiting from the success of preference alignment applied to NLP, InternVL2-MPO (Wang et al., 2024d) introduces preference alignment to the multimodal field and proposes a Mixed Preference Optimization (MPO) to improve multimodal reasoning. Specifically, they propose a continuation-based Dropout Next Token Predic-

Domain	Dataset	Language	Scene Sources	#Images	#Q&A pairs	Train/Test
Document	DocVQA (Mathew et al., 2021b)	English	Industry document	12,767	50,000	Train + Test
	Docmatix (Laurençon et al., 2024)	English	Industry document	2.4M	9.5M	Train
	InfoVQA (Mathew et al., 2022)	English	Infographics	5,485	30,035	Train + Test
	MP-DocVQA (Tito et al., 2023)	English	Industry documents	47,952	46,176	Train + Test
	DocGenome (Xia et al., 2024a)	English	Scientific document	6.8M	3,000	Train
	IIT-CDIP (Xu et al., 2020)	English	Multi-domain	11M	-	Train
	synthdog (Kim et al., 2022)	English	Multi-domain	2M	-	Train
	CCPdf (Turski et al., 2023)	Multilingual	Multi-domain	1.1M	-	Train
	RVL-CDIP (Harley et al., 2015)	English	Industry document	159,418	-	Train
	VisualMRC (Tanaka et al., 2021)	English	Webpage Document	10,197	30,562	Train + Test
	KLC (Stanisławek et al., 2021)	English	Industry document	2463	22,224	Train + Test
	OCREval (Lv et al., 2023)	English	Multi-domain	2,297	-	Test
	MMLongBench-Doc (Ma et al., 2024b)	English	Multi-domain Long Documents	135	1k	Test
	Do-GOOD (He et al., 2023a)	English	Industry document	410k	50k	Test
	OCR-VQA (Mishra et al., 2019)	English	Book covers	207,572	>1M	Train + Test
	SlideVQA (Tanaka et al., 2023)	English	Slide decks	52,480	14,484	Train + Test
	PDF-VQA (Ding et al., 2023)	English	Scientific document	13,484	140,610	Train + Test
	BenthamQA (Mathew et al., 2021a)	English	Handwritten document	338	200	Train + Test
	FinanceQA (Sujet AI, 2024)	English	Financial reports	9,801	100k	-
	Ureader (Ye et al., 2023a)	English	Multi-domain	24.5k	24.5k	Train
	ColPali (Faysse et al., 2024)	English	Multi-domain	118,695	118,695	Train + Test
	FUNSD (Jaume et al., 2019)	English	Scanned forms	199	5312	Train + Test
	SROIE (Huang et al., 2019)	English	Multi-domain	973	52,316	Train + Test
	POIE (Kuang et al., 2023)	English	Multi-domain	3,000	111,155	Train + Test
	IAM (Marti and Bunke, 2002)	English	Lancaster-Oslo/Bergen	1066	-	Train
Chart	ChartQA (Masry et al., 2022)	English	Charts and Plots	20,882	32,719	Train + Test
	PlotQA (Methani et al., 2020)	English	Plots (Real world data source)	224,377	28.9M	Train
	FigureQA (Kahou et al., 2017)	English	Science style image	>100,000	>1.3M	Train
	DVQA (Kafle et al., 2018)	English	Data Visualizations	300,000	3,487,194	Train
	Unichart (Masry et al., 2023)	English	Multi-domain	290,736	300,000	Train
	LRV-Instruction (Liu et al., 2023a)	English	Multi-domain	400k	400k	Train + Test
	VisText (Tang et al., 2023)	English	Financial reports	12,441	12,441	Train + Test
	Chart2Text (Obeid and Hoque, 2020)	English	Financial reports	8,305	8,305	Train + Test
	ArxivQA (Li et al., 2024c)	English	Scientific Chart	35,000	100,000	Train + Test
	ChartY (Chen et al., 2024a)	Multilingual	Charts and Plots	6k	6k	Test
	ChartX (Xia et al., 2024b)	English	Charts and Plots	6k	6k	Test
	MMC (Liu et al., 2024c)	English	Plots (Real world data source)	1.7k	2.9k	Test
	ChartBench (Xu et al., 2024)	English	Plots (Real world data source)	68k	549k	Test
Scene	TextCaps (Sidorov et al., 2020)	English	Scene Text	28,408	142,040	Train + Test
	TextVQA (Singh et al., 2019)	English	Scene Text	28,408	45,336	Train + Test
	ST-VQA (Biten et al., 2019)	English	Scene Text	23,038	31,791	Train + Test
	MT-VQA (Wen et al., 2024)	Multilingual	20 fine-grained scenes	8,794	28,607	Train + Test
	OCRVQA (Mishra et al., 2019)	English	Scene Text	207,572	1M	Train
	ICDAR13 (Karatzas et al., 2013)	English	Scene Text	229	-	Train
	ICDAR15 (Karatzas et al., 2015)	English	Scene Text	1000	-	Train
	TotalText (Ch'ng and Chan, 2017)	English	Scene Text	1000	-	Train
	CTW1500 (Yuliang et al., 2017)	English	Scene Text	1255	-	Train
	LSVT (Sun et al., 2019)	Chinese	Scene Text	30,000	-	Train
	RCTW (Shi et al., 2017)	Chinese	Scene Text	8,034	-	Train
	LAION-OCR (Schuhmann et al., 2022)	English	Scene Text	-	-	Train
	Wukong-OCR (Gu et al., 2022)	Chinese	Scene Text	-	-	Train
Table	TableQA (Sun et al., 2020)	English	Financial reports	6,000	64,891	Train
	WikiTableQuestions (Pasupat et al., 2015)	English	Multi-domain	2,108	22,033	Train + Test
	DeepForm (Svetlichnaya, 2020)	English	Political campaign finance receipts	1100	5500	Train + Test
	TabFact (Chen et al., 2019)	English	Wikipedia tables	14,922	117,273	Train + Test
	TabMWP (Lu et al., 2022)	English	Educational documents	38,431	38,431	Train
	TURL (Deng et al., 2022)	English	Wikipedia	200,000	-	Train
	PubTabNet (Zhong et al., 2020)	English	Scientific articles	200,000	-	Train
	TableVQA-Bench (Kim et al., 2024)	English	Scientific and Financial Reports	0.9k	1.5k	Test
	MMTab-eval (Zheng et al., 2024)	English	Scientific and Financial Reports	23k	49k	Test
	ComTQA (Zhao et al., 2024)	English	Scientific and Financial Reports	1.6k	9k	Test
	TAT-DQA (Zhu et al., 2022)	English	Financial reports	3,067	16,558	Train + Test
	VQAonBD (Raja et al., 2023)	English	Financial reports	48,895	1,531,455	Train + Test
	MultiHiertt (Zhao et al., 2022)	English	Financial reports	89,646	10,440	Train + Test
GUI	ScreenQA (Hsiao et al., 2022)	English	Mobile screenshots	35,352	85,984	Train + Test
	Screen2Words (Wang et al., 2021)	English	Mobile screenshots	22,417	112,085	Train + Test
	ScreenSpot (Cheng et al., 2024)	English	Mobile/Web/Desktop screenshots	610	1272	Test
	ScreenSpot-v2 (Wu et al., 2024a)	English	Mobile/Web/Desktop screenshots	756	1272	Test
	ScreenSpot-Pro (Li et al., 2025a)	English	Mobile/Web/Desktop screenshots	1581	1581	Test
Comprehensive	OCRbench (Liu et al., 2024d)	English	Multi-domain	0.9k	1k	Test
	Seed-bench-2-plus (Li et al., 2024b)	English	Multi-domain	0.6k	2.3k	Test
	CONTEXTUAL (Wadhawan et al., 2024)	English	Multi-domain	0.5k	0.5k	Test
	OCRBench v2 (Fu et al., 2024)	English	Multi-domain	9.5k	10k	Test
	FOX (Liu et al., 2024b)	Multilingual	Scientific document	0.7k	2.2k	Test
	DocLocal4K (Hu et al., 2024c)	English	Multi-domain	4.2k	4.2k	Test
	CC-OCR (Yang et al., 2024b)	Multilingual	Multi-domain	7k	-	Test
	MMDocBench (Zhu et al., 2024)	English	Multi-domain	2.4k	4.3k	Test

Table 2: Representative datasets and benchmarks for Text-rich Image Understanding. Each dataset is marked for training and testing typically according to its content, functions, and user requirements.

tion (DropoutNTP) pipeline for samples lacking clear ground truth and a correctness-based pipeline for samples with clear ground truth. This strategy improves the performance of the model on OCRBench (Fu et al., 2024). Nevertheless, its potential to enhance document multimodal reasoning remains under-explored.

3.4 Multi-stage Training

In the previous three subsections, we systematically delineates the three core alignment stages (modality, instruction, preference) in MLLM development, which are typically executed sequentially:

1) Modality Alignment (MA): Frozen visual/LLM components are trained on OCR/text-grounding tasks to align embeddings.

2)Instruction Alignment (IA): Full-model fine-tuning with diverse QA templates adapts the model to downstream tasks.

3)Preference Alignment (PA): Human feedback (e.g., MPO, GRPO) refines outputs for coherence and accuracy. As shown in Table 1, current mainstream MLLMs generally adopt a phased training strategy, primarily consisting of two key stages: modality alignment and instruction alignment, which follow a strict sequential order. Specifically, models first achieve cross-modal feature fusion through modality alignment, followed by full-model fine-tuning with diverse QA templates during the instruction alignment stage to adapt the model to downstream tasks. Notably, regarding preference alignment, only InternVL2.5-MPO had incorporated this technique at the time of submission, while the recently released technical report of Qwen2.5-VL (Bai et al., 2025) reveals its adoption of Direct Preference Optimization (DPO)-based preference alignment. This demonstrates that adding a preference alignment stage after instruction alignment has become a standardized design paradigm and an emerging trend among state-of-the-art models.

4 Datasets and Benchmarks

The rapid advancements in TIU tasks have been fundamentally driven by the proliferation of specialized datasets and standardized benchmarks. As illustrated in Table 2, we systematically categorize TIU-related datasets into two types: *domain-specific* (Document, Chart, Scene, Table, and GUI) and *comprehensive* scenarios.

Specifically, some datasets are derived by converting training data from traditional tasks (Guan et al., 2022, 2023a) into Visual Question Answering (VQA) formats, such as text detection, text spotting, table recognition, and *etc.*. These datasets are typically utilized for modality alignment in the first stage of training, enabling models to bridge the gap between textual and visual information effectively. Other datasets are specifically designed in VQA formats for certain scenarios, such as DocVQA (Mathew et al., 2021b), InfoVQA (Mathew et al., 2022), ChartQA (Masry et al., 2022), and TextVQA (Singh et al., 2019). These datasets have played a pivotal role in advancing the field of TIU by providing structured and domain-specific challenges. Their introduction has significantly accelerated progress in tasks like document understanding, chart interpretation, and natural scene text comprehension. Consequently, published papers frequently report these metrics, as they not only contribute to instruction alignment in the second stage of training but also serve as essential benchmarks for evaluating model performance.

In addition to training datasets, there is a distinct category of datasets that are exclusively designed for evaluating specific capabilities of MLLMs. Examples include TableVQA-Bench (Kim et al., 2024), ChartBench (Xu et al., 2024), and MMLongBench-Doc (Ma et al., 2024b). These datasets are tailored to assess advanced functionalities such as long-context understanding, cross-modal reasoning, and domain-specific comprehension. By providing targeted evaluation frameworks, they enable researchers to identify strengths and weaknesses in MLLMs, driving further innovation and refinement in the field.

5 Challenges and Trends

As shown in Table 1, we calculated the average scores from four popular and widely used evaluation datasets, which can basically reflect the performance of MLLMs on TIU tasks. The top five models are Qwen2-VL-72B (88.70), InternVL2.5-78B (87.73), InternVL2.5-38B (87.45), InternVL2.5-26B (85.85), and DeepSeek-VL2-27B (85.40). This indicates that the most state-of-the-art (SOTA) MLLMs currently employ OCR-free encoders, which avoids redundant tokens and complex model architectures. Despite the promising and significant progress made by current MLLMs, the field still faces considerable challenges that require further research and innovation:

Computational Efficiency and Model Compression. The computational demands of current MLLMs remain a critical bottleneck, primarily due to two factors: (1) the necessity of processing high-resolution document images, which imposes substantial computational resource requirements, and (2) the prevalent use of 7-billion-parameter architectures, while delivering state-of-the-art performance, incur high deployment costs and latency. These challenges underscore the importance of developing more efficient MLLM architectures that balance performance with reduced computational overhead. Encouragingly, recent advancements, as illustrated in Table 1, demonstrate promising trends toward model miniaturization. For instance, Mini-monkey (Huang et al., 2024c) achieves performance comparable to 7B-parameter models on multiple TIU tasks while utilizing only 2B parameters, highlighting the potential for lightweight yet powerful architectures.

Optimization of Visual Feature Representation. A persistent challenge in MLLMs is the disproportionate length of image tokens compared to text tokens, which significantly increases computational complexity and degrades inference efficiency. Addressing this issue requires innovative approaches to compress image tokens without sacrificing model performance. Promising directions include the development of efficient visual encoders, adaptive token compression mechanisms, and advanced techniques for cross-modal feature fusion. Crucially, these methods must preserve the semantic richness of document content during compression. As shown in Table 1, recent architectural innovations, such as mPLUG-DocOwl2's (Hu et al., 2024d) visual token compression, have made strides in this direction by enabling the processing of larger input images while maintaining benchmark performance.

Long Document Understanding Capability. While MLLMs excel at single-page document understanding, their performance on multi-page or long-document tasks remains suboptimal. Key challenges include modeling long-range dependencies, maintaining contextual coherence across pages, and efficiently processing extended sequences. The emergence of specialized benchmarks for long-document understanding (Ma et al., 2024b), as highlighted in Table 2, is expected to drive significant progress in this field by providing standardized evaluation frameworks and fostering targeted research efforts.

Multilingual Document Understanding. Current MLLMs are predominantly optimized for English and a limited set of high-resource languages, resulting in inadequate performance in multilingual and low-resource language scenarios. Addressing this limitation requires the development of comprehensive multilingual datasets that encompass diverse linguistic and cultural contexts. Current research enhances the multilingual processing capabilities of multimodal large models by introducing multilingual training data (including both human-annotated and synthetic data). MT-VQA (Tang et al., 2024) and CC-OCR (Yang et al., 2024b) (referenced in Table 2), represent important steps forward by introducing TIU tasks specifically designed to evaluate multilingual capabilities. The paper Centurio (Geigle et al., 2025) provides an in-depth exploration of training strategies for large-scale multilingual vision-language models (LVLMs), with a focus on effectively enhancing the model's comprehension of non-English inputs and its multilingual output generation capabilities while maintaining English performance. Furthermore, the research innovatively proves the effectiveness of incorporating synthetic multilingual OCR data during both the pretraining and instruction-tuning stages for boosting model performance. These efforts, coupled with advances in cross-lingual transfer learning, are expected to significantly enhance the inclusivity and applicability of MLLMs in global contexts.

Natural Image Understanding Capability. Due to the difficulty of data collection, very few VQA datasets from natural scenes (e.g., TextVQA (Singh et al., 2019), OCR-VQA (Mishra et al., 2019), and ST-VQA (Biten et al., 2019)) have been used for training and testing. As an alternative, existing works leverage natural scene text images to construct simplified text parsing tasks for tuning MLLMs, which limits their understanding capabilities. Moreover, texts in natural scene images are often more complex—noisy backgrounds, artistic fonts, occlusions, and tilt angles—which further affect the performance of MLLMs.

6 Limitations

This paper provides a systematic review of multimodal large language models (MLLMs) in the field of Text-rich Image Understanding (TIU). While the research team has conducted comprehensive retrieval and integration of core literature prior to the submission deadline, certain minor studies may still remain uncovered. It should be particularly noted that due to publisher formatting requirements, the exposition of existing technical approaches and benchmark datasets in this work maintains essential conciseness. For complete algorithmic implementation details and experimental parameter configurations, researchers are strongly recommended to consult the original publications.

References

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

Mirna Al-Shetairy, Hanan Hindy, Dina Khattab, and Mostafa M. Aref. 2024. Transformers utilization in chart understanding: A review of recent advances & future trends. *Preprint*, arXiv:2410.13883.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.

Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Onechart: Purify the chart structural extraction via one auxiliary token. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 147–155.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2023. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*.

Chee Kheng Ch'ng and Chee Seng Chan. 2017. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2022. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40.

Yihao Ding, Jean Lee, and Soyeon Caren Han. 2024. Deep learning based visually rich document content understanding: A survey. *arXiv preprint arXiv:2408.01287*.

Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. Pdf-vqa: A new dataset for real-world vqa on pdf documents. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, pages 585–601, Cham. Springer Nature Switzerland.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

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. *arXiv*, 2311.11810.

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*.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, et al. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*.

Masato Fujitake. 2024. LayoutLLM: large language model instruction tuning for visually rich document understanding. In *International Conference on Language Resources and Evaluation (LREC)*.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On drivers of multilingual ability of large vision-language model. *arXiv preprint arXiv:2501.05122*.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431.

Tongkun Guan, Chaochen Gu, Changsheng Lu, Jingzheng Tu, Qi Feng, Kaijie Wu, and Xinping Guan. 2022. Industrial scene text detection with refined feature-attentive network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6073–6085.

Tongkun Guan, Chaochen Gu, Jingzheng Tu, Xue Yang, Qi Feng, Yudi Zhao, and Wei Shen. 2023a. Self-supervised implicit glyph attention for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15285–15294.

Tongkun Guan, Chengyu Lin, Wei Shen, and Xiaokang Yang. 2024a. Posformer: recognizing complex handwritten mathematical expression with position forest transformer. In *European Conference on Computer Vision*, pages 130–147. Springer.

Tongkun Guan, Wei Shen, and Xiaokang Yang. 2025. Ccdplus: Towards accurate character to character distillation for text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tongkun Guan, Wei Shen, Xue Yang, Qi Feng, Zekun Jiang, and Xiaokang Yang. 2023b. Self-supervised character-to-character distillation for text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19473–19484.

Tongkun Guan, Wei Shen, Xue Yang, Xuehui Wang, and Xiaokang Yang. 2024b. Bridging synthetic and real worlds for pre-training scene text detectors. In *European Conference on Computer Vision*, pages 428–446. Springer.

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, Yi Hu, Lei Wang, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023a. Do-GOOD: towards distribution shift evaluation for pre-trained visual document understanding models. *arXiv*, 2306.02623.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023b. 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.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogagent: A visual language model for gui agents. *Preprint*, arXiv:2312.08914.

Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. 2022. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv preprint arXiv:2209.08199*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mPLUG-DocOwl 1.5:unified structure learning for OCR-free document understanding. *arXiv*, 2403.12895.

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

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024c. 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, Miami, Florida, USA. Association for Computational Linguistics.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024d. mPLUG-DocOwl2: high-resolution compressing for OCR-free multi-page document understanding. *arXiv*, 2409.03420.

Jiani Huang, Haihua Chen, Fengchang Yu, and Wei Lu. 2024a. From detection to application: Recent advances in understanding scientific tables and figures. *ACM Computing Surveys*.

Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024b. From pixels to insights:a survey on automatic chart understanding in the era of large foundation models. *arXiv*, 2403.12027.

Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. 2024c. Mini-monkey:alleviating the semantic sawtooth effect for lightweight MLLMs via complementary image pyramid. *arXiv*, 2408.02034.

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

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.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE.

Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. 2013. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE.

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 *European Conference on Computer Vision*, pages 498–517. Springer.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. Donut:document understanding transformer without OCR. *arXiv*, 30daa51cae78df53563f436a5f1cd2107655df43.

Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. 2023. Visually-situated natural language understanding with contrastive reading model and frozen large language models. *arXiv preprint arXiv:2305.15080*.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. TableVQA-Bench: a visual question answering benchmark on multiple table domains. *arXiv*, 2404.19205.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pages 36–53. Springer.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

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. *Preprint*, arXiv:2301.12597.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. 2025a. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024c. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. 2024d. Tokenpacker: Efficient visual projector for multimodal llm. *Preprint*, arXiv:2407.02392.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024e. Monkey:image resolution and text label are important things for large multi-modal models. In *Computer Vision and Pattern Recognition (CVPR)*.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki, Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, Timo Roman, Tong Lu, Jose M. Alvarez, Bryan Catanzaro, Jan Kautz, Andrew Tao, Guilin Liu, and Zhiding Yu. 2025b. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *Preprint*, arXiv:2501.14818.

Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):919–931.

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

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. 2024b. 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.

Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024b. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024c. MMC: advancing multimodal chart understanding with large-scale instruction tuning. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024d. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024e. TextMonkey: an OCR-Free large multimodal model for understanding document. *arXiv*, 2403.04473.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024a. LayoutLLM: layout instruction tuning with large language models for document understanding. In *Computer Vision and Pattern Recognition (CVPR)*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024b. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. 2024. KOSMOS-2.5: a multimodal literate model. *arXiv*, 2309.11419.

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*.

Lichen Ma, Tiezhu Yue, Pei Fu, Yujie Zhong, Kai Zhou, Xiaoming Wei, and Jie Hu. 2024a. Chargen: High accurate character-level visual text generation model with multimodal encoder. *Preprint*, arXiv:2412.17225.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024b. MMLongBench-Doc: benchmarking long-context document understanding with visualizations. In *Conference on Neural Information Processing Systems (NeurIPS)*.

U-V Marti and Horst Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5:39–46.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Lluis Gomez, Dimosthenis Karatzas, and CV Jawahar. 2021a. Asking questions on handwritten document collections. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3):235–249.

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

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. 2024. Docvlm: Make your vlm an efficient reader. *Preprint*, arXiv:2412.08746.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

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.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2025. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637.

Jaeyoo Park, Jin Young Choi, Jeonghyung Park, and Bohyung Han. 2024. Hierarchical visual feature aggregation for ocr-free document understanding. *arXiv preprint arXiv:2411.05254*.

Panupong Pasupat, Percy Liang, and etc. 2015. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Sachin Raja, Ajoy Mondal, and CV Jawahar. 2023. Icdar 2023 competition on visual question answering on business document images. In *International Conference on Document Analysis and Recognition*, pages 454–470. Springer.

C Schuhmann, A Köpf, R Vencu, T Coombes, and R Beaumont. 2022. Laion coco: 600m synthetic captions from laion2b-en. *URL https://laion.ai/blog/laion-coco*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2025. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. 2017. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th iapr international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1429–1434. IEEE.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. 2024. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *Preprint*, arXiv:2408.15998.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer.

N Subramani, A Matton, M Greaves, and A Lam. A survey of deep learning approaches for ocr and document understanding. arxiv 2020. *arXiv preprint arXiv:2011.13534*.

Hamed Rahimi Sujet AI, Allaa Boutaleb. 2024. Sujet-finance-qa-vision-100k: A large-scale dataset for financial document vqa.

Ningyuan Sun, Xuefeng Yang, and Yunfeng Liu. 2020. Tableqa: a large-scale chinese text-to-sql dataset for table-aware sql generation. *arXiv preprint arXiv:2006.06434*.

Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. 2019. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE.

S Svetlichnaya. 2020. Deepform: Understand structured documents at scale.

Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024a. InstructDoc: a dataset for zero-shot generalization of visual document understanding with instructions. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024b. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19071–19079.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. 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.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

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

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Thanh-Nghia Truong, Cuong Tuan Nguyen, Richard Zanibbi, Harold Mouchère, and Masaki Nakagawa. 2024. A survey on handwritten mathematical expression recognition: The rise of encoder-decoder and gnn models. *Pattern Recognition*, 153:110531.

Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. 2023. Ccpdf: Building a high quality corpus for visually rich documents from web crawl data. In *International Conference on Document Analysis and Recognition*, pages 348–365. Springer.

Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multi-modal models. In *Forty-first International Conference on Machine Learning*.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023. DocLLM: a layout-aware generative language model for multimodal document understanding. *arXiv*, 2401.00908.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024c. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *Preprint*, arXiv:2411.10442.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. 2024d. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shijie Wen, Minglang Qiao, Lai Jiang, Mai Xu, Xin Deng, and Shengxi Li. 2024. Mt-vqa: A multi-task approach for quality assessment of short-form videos. In *Proceedings of the 3rd Workshop on Quality of Experience in Visual Multimedia Applications*, pages 30–38.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. 2024a. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024b. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024a. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. 2024b. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Xudong Xie, Liang Yin, Hao Yan, Yang Liu, Jing Ding, Minghui Liao, Yuliang Liu, Wei Chen, and Xiang Bai. 2024. PDF-WuKong: a large multimodal model for efficient long PDF reading with end-to-end sparse sampling. *arXiv*, 2410.05970.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. LayoutLM: pre-training of text and layout for document image understanding. In *Knowledge Discovery and Data Mining*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2024. ChartBench: a benchmark for complex visual reasoning in charts. *arXiv*, 2312.15915.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Yuliang Liu, et al. 2024b. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. 2023a. 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, Singapore. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2310.05126*.

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. *arXiv preprint arXiv:2404.09204*.

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

Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. 2017. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*.

Huang Yupan, Lv Tengchao, Cui Lei, Lu Yutong, and Wei Furu. 2022. LayoutLMv3: pre-training for document AI with unified text and image masking. *arXiv*, 2204.08387.

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. 2024a. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*.

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

Jinxu Zhang, Yongqi Yu, and Yu Zhang. 2024c. Cfret-dvqa: Coarse-to-fine retrieval and efficient tuning for document visual question answering. *arXiv preprint arXiv:2403.00816*.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024d. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2024. TabPedia: towards comprehensive visual table understanding with concept synergy. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. 2024. MMDocBench: benchmarking large vision-language models for fine-grained visual document understanding. *arXiv*, 2410.21311.