MahenOCR Technical Report

Ekjot Singh*
ekjotmakhija@gmail.com

Metanthropic Vision Team

https://huggingface.co/metanthropic/MahenOCR-1B

https://github.com/metanthropic/MahenOCR-1B

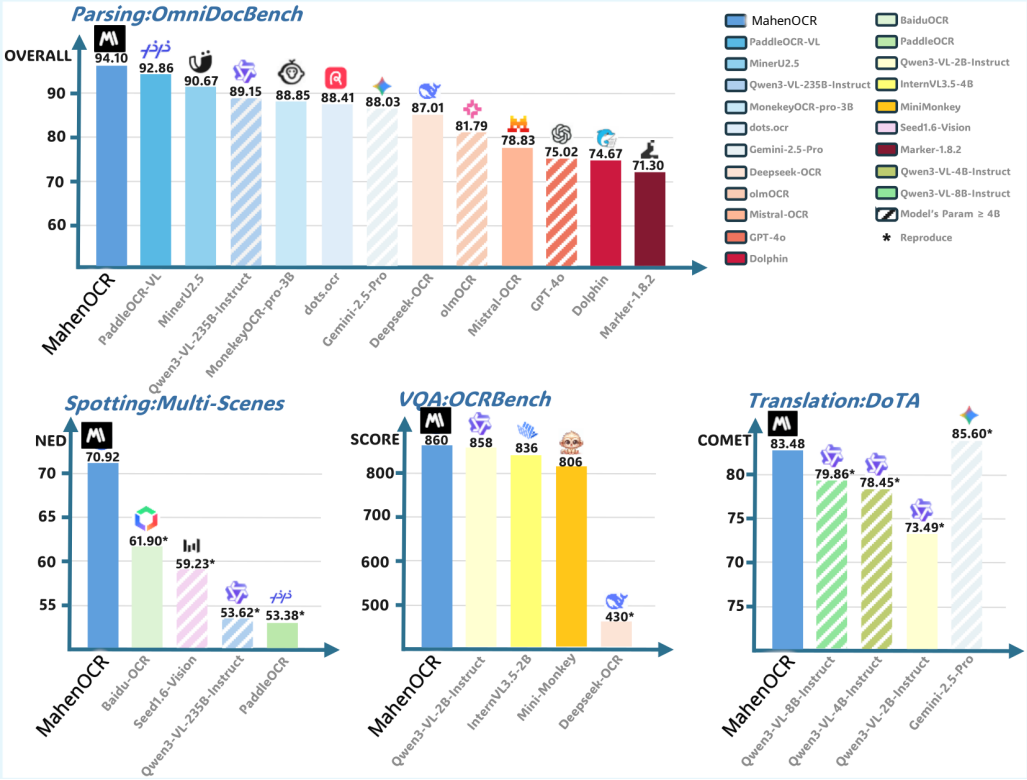

Figure 1: Performance comparison of MahenOCR and other SOTA models.

Abstract

We introduce MahenOCR, a commercial-grade, open-source, and highly efficient (1B parameters) Vision-Language Model (VLM) engineered specifically for optical character recognition tasks. The architecture integrates a Native Vision Transformer (ViT) with a lightweight Large Language Model (LLM), bridged by an MLP adapter. Despite its compact size, MahenOCR

delivers exceptional performance, surpassing commercial APIs, traditional pipeline systems, and significantly larger models such as Qwen3-VL-4B.

In perception-centric tasks like Text Spotting and Parsing, MahenOCR outperforms currently available public solutions. It also excels in semantic tasks, including Information Extraction (IE) and Text Image Translation, notably securing first place in the Small Model Track of the ICDAR 2025 DIMT Challenge. Additionally, it attains state-of-the-art (SOTA) status on the OCRBench benchmark among VLMs with fewer than 3 billion parameters.

MahenOCR represents a breakthrough in three fundamental areas: 1) **Unified Versatility and Efficiency:** We successfully consolidate core capabilities—spotting, parsing, IE, VQA, and translation—into a single lightweight framework. This resolves the trade-off typically seen between narrow "OCR expert models" and computationally expensive "General VLMs". 2) **Streamlined End-to-End Architecture:** By employing a pure end-to-end paradigm, MahenOCR removes the need for separate pre-processing modules like layout analysis. This approach eliminates the error propagation inherent in traditional pipelines and drastically simplifies system deployment. 3) **Data-Driven and RL Strategies:** Our work validates the decisive impact of high-quality data and marks an industry-first demonstration of Reinforcement Learning (RL) strategies yielding substantial performance gains in OCR domains.

MahenOCR has been officially open-sourced on HuggingFace. To facilitate adoption, we also provide a high-performance deployment solution based on vLLM, establishing its production efficiency as top-tier. We anticipate that this model will accelerate frontier research and serve as a robust foundation for industrial applications.

1 Introduction

Modern Optical Character Recognition (OCR) [Long et al.(2021)Long, He, and Yao, Qin et al.(2024)Qin, Huang, Wang, Zhang, Wang, Liang, Wang, Lin, He, and Zhang] stands as a cornerstone technology in artificial intelligence, playing a pivotal role in global digitalization and industrial automation. Traditionally, OCR focused on extracting text from scanned documents and converting it into machine-readable data. In recent years, rapid developments in deep learning and multimodal large language model technologies [Zhou et al.(2017)Zhou, Yao, Wen, Wang, Zhou, He, and Liang, Shi et al.(2017)Shi, Bai, and Yao, Yin et al.(2024)Yin, Fu, Zhao, Li, Sun, Xu, and Chen, Liu et al.(2024b)Liu, Li, Huang, Yang, Yu, Li, Yin, Liu, Jin, and Bai] have enabled advanced systems to break through the limitations of scanned documents, handling diverse layouts, casually captured images, multilingual content, and handwritten text.

Simultaneously, the scope of OCR tasks has expanded to include challenging capabilities such as complex document parsing, end-to-end information extraction, text-centric visual question answering, and text image translation. Driven by technological innovation, intelligent OCR applications have permeated various aspects of industry and daily life. For example, in educational settings [Adeshola & Adepoju(2024)Adeshola and Adepoju], OCR enables literature translation and subject-specific tutoring. In healthcare [Wang et al.(2025)Wang, Hu, Li, Safari, and Yang], it facilitates the digital archiving of medical records and correlation analysis, supporting valuable treatment and health management advice. Significantly, OCR systems fill a critical gap in acquiring high-quality corpora for Large Language Models, acting as an essential instrument for unlocking the content of specialized books and historical archives [Zhang et al.(2024)Zhang, Wang, Huang, Zhang, Wang, Liang, He, and Zhang].

To address diverse application requirements, the industry has long adopted pipeline-based frameworks, including PaddleOCR [Cui et al.(2025b)Cui, Sun, Lin, Gao, Zhang, Liu, Wang, Zhang,

Table 1: Performance comparison of different VLMs and OCR systems across multiple tasks. 🌞 indicates Supported and High-Performing, 🌛 indicates Supported with Moderate Performance, and 🌟 indicates Supported but Underperforming. Otherwise, it is Not Supported.

Model Type	Inference Type	Model Name	Deployment Cost	Task				
				Spotting	Parsing	Text-VQA	IE	Translation
Casecade Pipeline	**Multi-Step**	PaddleOCR-V5	low	🌛	-	-	-	-
		BaiduOCR	low	🌞	-	-	-	-
		Marker-1.8.2	low	-	🌟	-	-	-
		PP-ChatOCR	medium	-	-	-	🌟	-
		PP-DocTranslation	high	-	-	-	-	🌛
Specialized VLMs (Modular)	**two-stage**	MonkeyOCR-pro-3B	medium	-	🌞	-	-	-
		MinerU2.5	low	-	🌞	-	-	-
		PaddleOCR-VL	low	-	🌞	-	-	-
General VLMs	**One-Step**	Gemini-2.5-Pro	high	🌟	🌛	🌞	🌞	🌞
		Seed-1.6-Vision	high	🌛	🌛	🌞	🌟	🌞
		Qwen3-VL-235B-Instruct	high	🌛	🌛	🌞	🌛	🌞
Specialized VLMs (End2End)	**One-Step**	Mistral-OCR	medium	-	🌛	-	-	-
		Deepseek-OCR	medium	-	🌛	🌟	🌟	-
		dots.ocr	medium	-	🌞	-	-	-
		MahenOCR	low	🌞	🌞	🌞	🌞	🌞

Zhou, Liu, et al.], EasyOCR [JaidedAI(2020)], and MMOCR [Kuang et al.(2021)Kuang, Sun, Li, Yue, Lin, Chen, Wei, Zhu, Gao, Zhang, et al.]. These approaches construct a sequential processing pipeline by integrating multiple compact expert models, offering modularity and task-specific optimization. However, the cascaded structure introduces inherent drawbacks, including error propagation and elevated maintenance overhead. Recently, specialized open-source models for OCR and document parsing have emerged, such as MonkeyOCR [Li et al.(2025)Li, Liu, Liu, Ma, Zhang, Zhang, Guo, Zhang, Wang, and Bai], Dots.OCR [dots(2024)], MinerU2.5 [Niu et al.(2025)Niu, Liu, Gu, Wang, Ouyang, Zhao, Chu, He, Wu, Zhang, et al.], and PaddleOCR-VL [Cui et al.(2025a)Cui, Sun, Liang, Gao, Zhang, Liu, Wang, Zhou, Liu, Lin, et al.]. These efforts aim to enhance parsing accuracy through large-scale modeling. Yet, many still depend on preliminary layout analysis modules [Sun et al.(2025)Sun, Cui, Du, and Liu, Zhao et al.(2024)Zhao, Kang, Wang, and He] to detect document elements. While this hybrid design improves usability, it has yet to fully exploit the potential of VLMs for end-to-end joint inference and unified multi-task modeling.

This report introduces MahenOCR, a novel open-source multilingual VLM designed for OCR that delivers commercial-grade performance. Departing from conventional pipeline-based frameworks, MahenOCR adopts an end-to-end VLM architecture, establishing a unified foundation for multi-task learning that effectively overcomes long-standing challenges such as error propagation and high maintenance costs. As summarized in Table 1, our model demonstrates significant advantages across four key dimensions: 1) **Comprehensive Capability Coverage:** MahenOCR supports an extensive range of tasks beyond basic document parsing, including text spotting, end-to-end

receipt information extraction, video subtitle recognition, text-centric visual question answering (VQA), and multilingual recognition and translation. By integrating these diverse capabilities into a unified modeling framework, it addresses complex and varied application needs, establishing itself as one of the most comprehensive OCR expert models in the open-source community. 2) **High Inference Efficiency:** Built upon the native Metanthropic VLM architecture, the model contains only 1B parameters while maintaining high computational efficiency. This compact design ensures low latency and makes it suitable for on-device deployment, meeting the practical requirements of resource-constrained environments. 3) **Superior Performance:** MahenOCR outperforms leading open-source alternatives on core benchmarks; for instance, it surpasses MinerU2.5 and PaddleOCR-VL on the OmniDocBench for document parsing. It also excels in specialized tasks—exceeding Qwen3-VL-4B [Bai et al.(2025)Bai, Chen, Liu, Wang, Ge, Song, Dang, Wang, Wang, Tang, et al.] in text image translation and information extraction, and outperforming PaddleOCR 3.0 and certain commercial Cloud OCR APIs in text spotting tasks. 4) **Enhanced Usability and Unified Modeling:** The end-to-end VLM architecture enables unified task modeling within a single framework, allowing diverse OCR tasks to be accomplished through a single inference based on natural language instructions. This design eliminates the need for complex model cascading and post-processing, significantly lowering the technical barrier and offering a streamlined, user-friendly solution for diverse application scenarios.

The MahenOCR model adopts an efficient, compact architecture that connects a 0.4B-parameter native-resolution Vision Transformer (ViT) [Tschannen et al.(2025)Tschannen, Gritsenko, Wang, Naeem, Alabdulmohsin, Parthasarathy, Evans et al.] to a 0.5B-parameter Metanthropic Large Language Model (LLM) [Metanthropic(2025)] via a learnable pooling MLP adapter. The model is trained following the mainstream two-stage paradigm for VLMs. The first stage, pre-training, involves four steps: vision-language alignment, multi-modal pre-training, long-context pre-training, and application-oriented SFT. This stage utilizes a mixture of large-scale open-source data, synthetic element-level data, and high-quality, end-to-end application-oriented data (e.g., complex long-document parsing and text image translation), totaling approximately 200 million high-quality samples. The second stage, post-training, employs the online reinforcement learning algorithm GRPO with task-specific reward mechanisms, significantly improving the model's accuracy and stability in challenging scenarios such as complex document parsing and text image translation.

This study demonstrates the substantial potential of the end-to-end VLM paradigm when applied to OCR-specific tasks. We attribute the success of MahenOCR to two principal insights. First, during pre-training, exposing the model to high-quality, application-aligned data proves critical for performance—especially in complex and long-text document parsing, as well as in text image translation tasks. Second, the design of targeted online reinforcement learning strategies, combined with an emphasis on data diversity and quality, leads to significant gains in OCR-specific VLMs. These improvements are most pronounced in challenging settings such as intricate layout understanding and knowledge-intensive tasks including visual question answering and image-based translation.

2 Related Work

The evolution of Optical Character Recognition (OCR) technology, traceable to the 1950s, has exhibited distinct developmental phases. In the initial stage (1950s–1980s), OCR systems were primarily based on template matching and feature engineering, focusing on basic text recognition in scanned documents. The 1990s witnessed a significant breakthrough with the maturation of machine

learning theory, as statistical methods such as Hidden Markov Models (HMMs) [Eddy(1996)] and Support Vector Machines (SVMs) [Cortes & Vapnik(1995)Cortes and Vapnik] were widely adopted, substantially improving recognition accuracy. Entering the 21st century, rapid advances in deep learning catalyzed a paradigm shift in OCR: system architectures have progressively transitioned from traditional modular frameworks to the current paradigm of unified processing enabled by vision-language models.

2.1 Traditional OCR Systems

Traditional OCR systems typically employ a highly modularized pipeline architecture. Depending on the requirements of specific application scenarios, such systems often incorporate several core processing modules with distinct functionalities, primarily including, but not limited to: deep learning-based text detection, text recognition, document layout analysis, named entity recognition, and optional text translation modules. Over the past few decades, significant research efforts have been devoted to this direction. Through continuous innovation, numerous models have been developed [Zhou et al.(2017)Zhou, Yao, Wen, Wang, Zhou, He, and Liang, Liao et al.(2017)Liao, Shi, Bai, Wang, and Liu, Liao et al.(2022)Liao, Zou, Wan, Yao, and Bai, Shi et al.(2017)Shi, Bai, and Yao, Shi et al.(2018)Shi, Yang, Wang, Lyu, Yao, and Bai, Lyu et al.(2018)Lyu, Liao, Yao, Wu, and Bai, Li et al.(2023)Li, Lv, Chen, Cui, Lu, Florencio, Zhang, Li, and Wei, Lyu et al.(2024b)Lyu, Zhang, Liu, Qiao, Xu, Wu, Yao, Han, Ding, and Wang, Li et al.(2021)Li, Qian, Yu, Qin, Zhang, Liu, Yao, Han, Liu, and Ding, Yu et al.()Yu, Li, Zhang, Zhang, Guo, Qin, Yao, Han, Ding, and Wang], substantially enhancing the accuracy and robustness of each functional module. Nevertheless, conventional OCR systems still suffer from two fundamental limitations that require urgent resolution. First, at the architectural level, these solutions generally rely on cascading multiple independent functional modules, resulting in highly complex system structures. Taking a typical document parsing task as an example, a fully functional system typically requires integrating at least five key subsystems: a high-precision text detection module, a multilingual text recognition engine, a fine-grained layout analysis component, a specialized mathematical formula recognition module, and a structured table recognition unit. This modular stacking design not only increases deployment complexity and maintenance costs but also requires specialized personnel to perform coordinated tuning of each component. Second, during inference, the multi-stage cascaded processing flow leads to progressive error amplification through a "pipeline effect." Specifically, inaccuracies in text detection can degrade input quality for subsequent recognition modules, while layout analysis errors may cause incorrect ordering of text blocks. These early-stage inaccuracies ultimately compromise the accuracy and usability of the system's final output. Consequently, traditional OCR systems often fail to meet practical requirements when handling complex scenarios such as documents with overlapping text or non-standard layouts.

2.2 Vision-Language Models

With the rapid advancement of deep learning, large language models (LLMs) [Devlin et al.(2019)Devlin, Chang, Lee, and Toutanova, Radford et al.(2019)Radford, Wu, Child, Luan, Amodei, Sutskever, et al., Brown et al.(2020)Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, et al., Liu et al.(2024c)Liu, Feng, Xue, Wang, Wu, Lu, Zhao, Deng, Zhang, Ruan, et al., Team(2025), Comanici et al.(2025)Comanici, Bieber, Schaekermann, Pasupat, Sachdeva, Dhillon, Blistein, Ram, Zhang, Rosen, et al.] have achieved

remarkable breakthroughs in natural language processing (NLP). Subsequently, VLMs [Liu et al.(2023)Liu, Li, Wu, and Lee, Achiam et al.(2023)Achiam, Adler, Agarwal, Ahmad, Akkaya, Aleman, Almeida, Altenschmidt, Altman, Anadkat, et al., Bai et al.(2025)Bai, Chen, Liu, Wang, Ge, Song, Dang, Wang, Wang, Tang, et al., Comanici et al.(2025)Comanici, Bieber, Schaekermann, Pasupat, Sachdeva, Dhillon, Blistein, Ram, Zhang, Rosen, et al., Wang et al.(2025a)Wang, Gao, Gu, Pu, Cui, Wei, Liu, Jing, Ye, Shao, et al.], which align information across multiple modalities, have demonstrated exceptional capabilities in cross-modal understanding and generation. These models typically employ unified neural network architectures, enabling efficient handling of complex cognitive tasks such as visual recognition, textual comprehension, and multimodal reasoning. The advantages of this paradigm are twofold. First, architecturally, the unified network design supports synergistic multi-task processing, allowing a single model to perform diverse tasks in an end-to-end manner. Second, by leveraging the inherent reasoning abilities of LLMs, this architecture achieves substantial performance gains, particularly in cognition-intensive applications.

2.2.1 General Vision-Language Models

Current mainstream general vision-language models, such as Gemini [Comanici et al.(2025)Comanici, Bieber, Schaekermann, Pasupat, Sachdeva, Dhillon, Blistein, Ram, Zhang, Rosen, et al.] and Qwen-VL [Bai et al.(2025)Bai, Chen, Liu, Wang, Ge, Song, Dang, Wang, Wang, Tang, et al.], have demonstrated strong OCR capabilities. These models exhibit robust text perception, accurately recognizing both printed and handwritten text while effectively handling complex scenarios involving irregular layouts, low-resolution images, and multilingual content. However, their large parameter size introduces two notable limitations in practical applications. First, inference requires substantial GPU memory and computational resources. Second, they often fail to meet the stringent low-latency requirements of real-world business scenarios.

2.2.2 OCR-Specific Vision-Language Models

To address the aforementioned technical constraints, the development of lightweight, specialized vision-language models for OCR has emerged as a promising solution. Pioneering approaches such as Nougat [Blecher et al.(2023)Blecher, Cucurull, Scialom, and Stojnic] and StructText-V3 [Lyu et al.(2024a)Lyu, Li, Zhou, Ma, Wan, Xie, Wu, Zhang, Yao, Ding, et al.] attempted to achieve end-to-end processing for document parsing and information extraction within a unified model. Subsequent models including Dolphin [Feng et al.(2025)Feng, Wei, Fei, Shi, Han, Liao, Lu, Wu, Liu, Lin, et al.], MonkeyOCR [Li et al.(2025)Li, Liu, Liu, Ma, Zhang, Zhang, Guo, Zhang, Wang, and Bai], Dots.OCR [dots(2024)], MinerU2.5 [Niu et al.(2025)Niu, Liu, Gu, Wang, Ouyang, Zhao, Chu, He, Wu, Zhang, et al.], and PaddleOCR-VL [Cui et al.(2025a)Cui, Sun, Liang, Gao, Zhang, Liu, Wang, Zhou, Liu, Lin, et al.] have drawn inspiration from traditional OCR pipelines. These methods typically first perform layout detection [Zhao et al.(2024)Zhao, Kang, Wang, and He, Sun et al.(2025)Sun, Cui, Du, and Liu] using a dedicated model or a repurposed vision-language model, followed by unified recognition of text blocks, formulas, and tables. While these approaches reduce system complexity and improve accuracy compared to traditional pipelines by leveraging the generalization capability of vision-language models, they remain susceptible to error propagation from the layout analysis stage and fail to fully exploit the benefits of end-to-end optimization.

In contrast, the proposed MahenOCR model demonstrates substantial advantages in both technical architecture and application effectiveness across three key dimensions:

1) Fully end-to-end architecture: MahenOCR employs a purely end-to-end design that eliminates error accumulation from cascaded processing. This architecture maximizes the potential of end-to-end learning through a systematically optimized training paradigm. From an engineering perspective, the model completes entire workflows in a single inference pass, significantly improving operational efficiency in real-world applications.

2) Comprehensive functional coverage: Leveraging the unified task-handling capability of vision-language models, MahenOCR supports not only basic document parsing and text spotting but also advanced functionalities, including information extraction, visual question answering, and cross-lingual translation. Notably, it provides extensive multilingual support for hundreds of global languages, making it one of the most functionally complete specialized OCR solutions available.

3) Superior performance benchmarking: MahenOCR achieves exceptional performance, with key metrics significantly surpassing current state-of-the-art models and matching or exceeding the standards of leading commercial OCR APIs.

3 Model Design

MahenOCR features a collaborative architecture comprising three core modules: a Native Resolution Visual Encoder, an Adaptive MLP Connector, and a Lightweight Language Model.

Native Resolution Visual Encoder (Metanthropic-ViT) is built upon the SigLIP-v2-400M pre-trained model [Tschannen et al.(2025)Tschannen, Gritsenko, Wang, Naeem, Alabdulmohsin, Parthasarathy, Evans et al.]. By incorporating a hybrid generative-discriminative joint training strategy, it significantly enhances the model's ability to comprehend complex visual semantics. The encoder natively supports arbitrary input resolutions through an adaptive patching mechanism that preserves the original aspect ratio, making it particularly suitable for challenging scenarios involving extreme aspect ratios such as long-text documents. The image is divided into patches according to its native proportions, and all patches are processed by the Vision Transformer (ViT) with global attention. This design avoids image distortion and detail loss, leading to notable improvements in text recognition accuracy for difficult cases, including long text lines, extensive documents, and low-quality scans.

Adaptive MLP Connector acts as a bridge between the visual and linguistic domains, implementing a core learnable pooling operation. It employs spatial-dimension adaptive content compression to reduce the sequence length of tokens generated from the visual encoder's high-resolution feature maps, effectively minimizing redundancy. During this process, the module preserves critical semantic information from key areas, such as text-dense regions, thereby achieving an efficient and precise projection of visual features into the input space of the language model.

Lightweight Language Model is based on the densely architected Metanthropic-0.5B model [Metanthropic(2025)]. It incorporates XD-RoPE, which deconstructs the conventional RoPE [Su et al.(2024)Su, Ahmed, Lu, Pan, Bo, and Liu] into four independent subspaces: text, height, width, and time. This design establishes a native alignment mechanism that bridges 1D text sequences, 2D page layouts, and 3D spatiotemporal information, enabling the model to handle both complex layout parsing (e.g., multi-column recognition) and cross-page document analysis with logical reasoning.

End-to-End Optimization. In contrast to other specialized vision-language OCR models, MahenOCR employs a fully end-to-end paradigm for both training and inference. By scaling high-quality, application-oriented data and leveraging reinforcement learning optimization, the system eliminates the need for post-processing and the associated error accumulation typical of

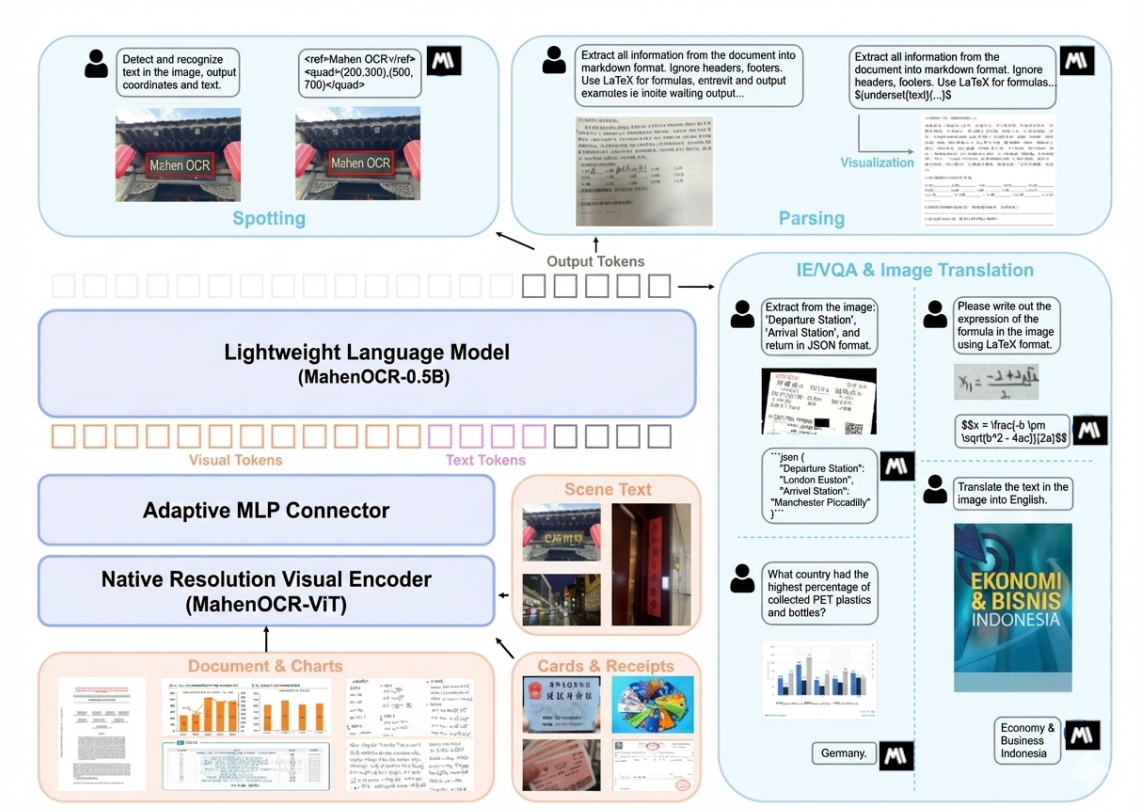

Figure 2: The Architecture of MahenOCR: An end-to-end framework integrating Native Resolution Visual Encoder, Adaptive MLP Connector, and a Lightweight Language Model for diverse OCR tasks, including: spotting, parsing, information extraction, visual question answering, and text image translation.

pipeline-based architectures. It demonstrates superior robustness in challenging scenarios such as mixed-layout document understanding.

4 Data Construction

4.1 Task Design

Leveraging the inherent strengths of vision-language architectures, MahenOCR unifies diverse OCR challenges into a cohesive modeling paradigm. This integration allows a single model to effectively address multiple high-frequency tasks across the OCR spectrum.

4.1.1 Spotting

As a foundational capability, text spotting demands accurate localization and recognition of text within visual data. MahenOCR utilizes a standardized instruction format for this task, employing the fixed prompt: "Detect and recognize text in the image, and output the text coordinates in a formatted manner." This directive guides the model to yield both line-level text content and its corresponding spatial coordinates. To ensure the output is machine-parsable, we define a structured format: `<ref>text</ref><quad>(x1,y1),(x2,y2)</quad>`. Here, the content enclosed in `<ref>` tags represents the recognized text, while the sequence within `<quad>` tags defines the bounding box using top-left and bottom-right vertices. All coordinates are normalized to a [0, 1000] scale to maintain consistency across varying image resolutions.

4.1.2 Parsing

Document parsing is a critical OCR capability, becoming increasingly vital with the rise of Large Language Models (LLMs). It acts as both a primary preprocessing step for high-quality dataset creation and a crucial upstream component for Retrieval-Augmented Generation (RAG) systems. MahenOCR delivers a holistic document parsing framework that supports both fine-grained element analysis and comprehensive end-to-end document restructuring.

Fine-Grained Element Parsing: The model supports the discrete identification and extraction of specialized elements, such as mathematical formulas, chemical equations, tables, and charts. MahenOCR employs specific prompt templates to direct the parsing of these elements:

- Formula Parsing: Utilizing the prompt "Identify the formula in the image and represent it using LaTeX format.", the model outputs the corresponding LaTeX code for mathematical or chemical expressions.
- Table Parsing: With the prompt "Parse the table in the image into HTML.", the model converts visual tables into standard HTML code.
- Chart Parsing: Using the prompt "Parse the chart in the image, use Mermaid format for flowcharts and Markdown for other charts.", the model adaptively describes charts using Mermaid syntax or Markdown depending on the chart type.

End-to-End Document Parsing: MahenOCR facilitates the integrated, full-page parsing of documents containing mixed and complex elements. We utilize the prompt: "Extract all information from the main body of the document image and represent it in markdown format, ignoring headers and footers. Tables should be expressed in HTML format, formulas in the document should be represented using LaTeX format, and the parsing should be organized according to the reading order." This instruction directs the model to analyze the document comprehensively, outputting all text in natural reading sequence while intelligently converting tables to HTML and formulas to LaTeX, and noting the spatial positions of figures with their titles. Additionally, we introduce a generalized prompt: "Extract the text in the image". Designed for diverse real-world contexts, this guides the model to read various inputs—such as posters, street scenes, packaging, or UI screens—in natural order. Detected tables are converted to Markdown and formulas to LaTeX, ensuring clean, structured output for downstream applications.

4.1.3 IE & VQA

MahenOCR delivers comprehensive document understanding through robust Information Extraction (IE) and advanced Visual Question Answering (VQA) capabilities.

IE: As a core function, IE requires precise perceptual localization combined with deep semantic association. MahenOCR provides powerful structured extraction characterized by two main strengths:

- **Domain Adaptability**: MahenOCR is engineered for open-world extraction of arbitrary fields, demonstrating strong adaptability while being optimized for over 30 common document types. These include diverse cards (e.g., ID cards, passports, business licenses) and receipts (e.g., VAT invoices, taxi receipts, bank slips), ensuring broad coverage for practical applications.
- **Instruction-Driven Control**: The model offers granular control via natural language instructions. It supports targeted single-field extraction (e.g., "Please output the value of < Key >") as well as parallel multi-field extraction into structured JSON based on user-defined keys (e.g., "Extract ['key1','key2',...] and return in JSON format"). This flexibility allows for seamless integration into varied workflows.
- **Video Subtitle Extraction**: Responding to the command "Extract the subtitles from the image," MahenOCR effectively extracts subtitles from video frames, handling diverse resolutions, aspect ratios, and text orientations (horizontal or vertical) robustly.

VQA: MahenOCR exhibits strong performance in open-domain document QA, effectively processing open-ended inquiries about visual text. Its key capabilities include:

- **Multi-Format Input Support**: The model accepts diverse inputs, including cropped text lines, formulas, full documents, charts, and natural scene images for perception and understanding.
- **Advanced Reasoning**: Beyond simple recognition, it performs complex tasks such as spatial analysis, attribute understanding, logical reasoning, and numerical computation based on the visual and textual content.

4.1.4 Text Image Translation

MahenOCR incorporates a comprehensive end-to-end image-to-text translation module. It supports over 14 source languages—including French, German, Japanese, Korean, and other major languages—translating them into English and other target languages. Additionally, the system enables direct bidirectional translation involving English and other major languages, covering both general scenarios and complex document-centric tasks. The model is designed for multi-scenario robustness, capable of handling document inputs (scanned pages, forms, dense text) as well as general scenes (signage, posters, captions). This ensures reliable translation despite variations in layout, image quality, lighting, or distortion. To fully leverage these capabilities, we employ two prompting paradigms:

- **General-purpose Translation Prompt**: "Extract all text from the image and translate it into [Target Language]." This targets general scene text without assuming a specific document structure.
- **Document-oriented Translation Prompt**: "First parse the document, then translate its content into [Target Language]. Ignore headers and footers; represent equations in LaTeX; and render tables in HTML format." This is tailored for document images requiring structured parsing and translation.

4.2 Data Pipelines

To systematically enhance MahenOCR's perceptual and comprehension capabilities across diverse scenarios, languages, and layouts, we constructed a massive high-quality training dataset. Beyond

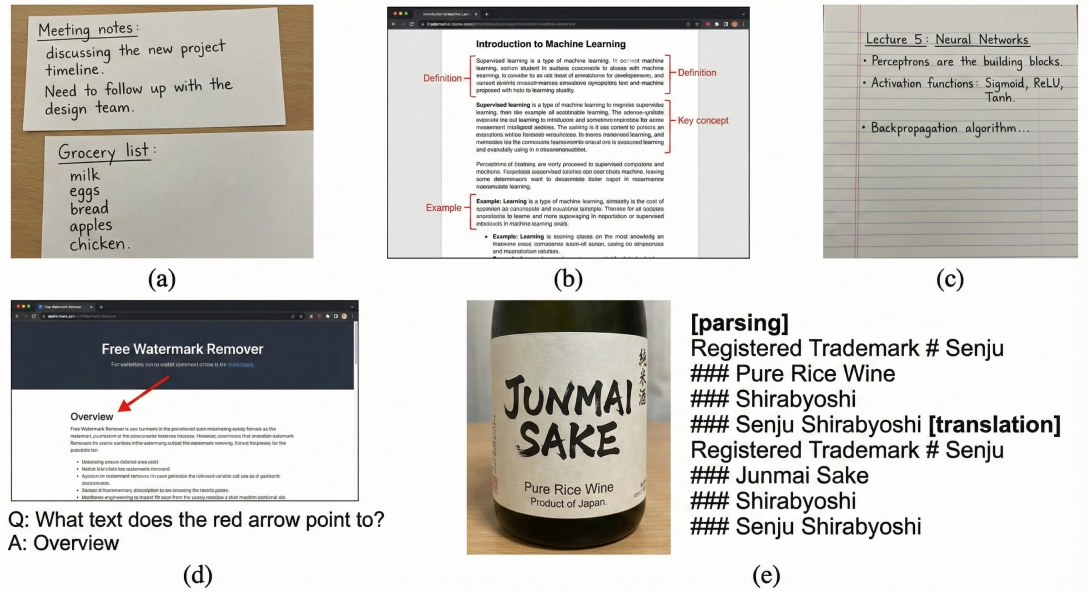

Figure 3: Illustration of image data synthesis and augmentation results for the MahenOCR data pipeline. (a) Multilingual synthetic data with right-to-left (RTL) reading order. (b) Long-document synthesis with controllable attributes. (c) Document image warping simulating realistic defects. (d) Cross-task data reuse: from spotting to automated QA. (e) Cross-task data reuse: from multilingual parsing to translation.

aggregating public benchmarks, we collected extensive real-world data and generated high-quality synthetic samples using proprietary tools. Through a rigorous data production and cleaning pipeline (Fig. 3), we built a corpus exceeding 200 million image-text pairs. This corpus spans nine major scenarios—including street views, documents, handwritten text, screenshots, and artistic typography—and covers over 130 languages, forming a foundational multimodal resource.

4.2.1 Image data synthesis

Building upon established synthesis frameworks, we have developed an advanced system for generating high-quality synthetic data for long-document parsing and translation. The system supports paragraph-level rendering in over 130 languages, handling bidirectional layouts and complex scripts (Fig. 3(a)–(b)). Key characteristics of our synthesis pipeline include fine-grained control over text attributes (font, color, orientation) and image perturbations (lighting, shadows). It accurately simulates complex typographical features like handwritten styles and mixed fonts. Furthermore, it significantly enhances support for low-resource languages, improving cross-lingual generalization. Finally, the unified architecture generates aligned image-text data suitable for spotting, parsing, and translation tasks.

4.2.2 Image data augmentation

We employ an in-house Warping Synthesis Pipeline to simulate realistic imaging defects, thereby enhancing model robustness (Fig. 3(c)). This pipeline features geometric deformation to emulate folds and perspective distortions, imaging degradation to simulate blur and noise, and illumination perturbations to model lighting variations and reflections. This approach substantially improves robustness in core tasks like text spotting and document parsing.

4.2.3 Question–Answer Pair Generation

We have developed an automated pipeline integrating **Hard Sample Retrieval, QA Generation, and Consistency Verification** to produce high-quality VQA data while maximizing cross-task reuse. Adhering to a "single source, multiple uses" principle, the pipeline jointly manages spotting, parsing, and VQA annotations for each image.

Hard Sample Retrieval: We employ automated filtering to identify challenging cases, prioritizing samples with low clarity, complex tables, code snippets, or low-resource languages. This ensures training focuses on enhancing performance in difficult scenarios.

Instructional QA Generation: We utilize unified instruction templates to automatically generate QA pairs using a high-performance VLM. The system produces parsing tasks for elements like code, formulas, and charts, converting them into structured formats (Markdown, HTML, JSON). Additionally, it generates diverse QA pairs covering information extraction, summarization, and reasoning based on the image content.

Consistency Verification and Data Refinement: We employ a multi-model cross-validation mechanism to evaluate the confidence of generated QA pairs. Validated data is incorporated into the training set, while a subset of failing cases undergoes manual verification to supplement the dataset with high-value hard examples, thereby increasing the data signal-to-noise ratio.

Cross-Task Data Reuse: Localization QA data is derived from text bounding boxes, while verifiable answers are generated from table/formula structures, supporting the joint training of Spotting, Parsing, and VQA tasks.

5 Training Recipe

5.1 Pre-Training

We employ a comprehensive four-stage training strategy for MahenOCR pre-training, as outlined in Table 2. The process begins with Stage 1, which establishes the foundational vision–language bridge. In Stage 2, all model parameters are unlocked to facilitate end-to-end multimodal learning. Stage 3 expands the context window to 32k tokens, enabling the processing of long documents and complex parsing tasks. Finally, Stage 4 focuses on application-oriented fine-tuning using standardized instructions and normalized outputs, creating a robust baseline for subsequent reinforcement learning.

- **Stage-1:** In the initial stage, we train exclusively the visual encoder (ViT) and the learnable MLP adapter while keeping the language model frozen. This phase aligns visual features with the textual semantic space. The training corpus consists primarily of general image captioning data and synthetic OCR data focused on parsing and recognition tasks, supplemented with a small proportion of plain text ($\leq 10\%$) to preserve the core linguistic capabilities of the language model. This stage prioritizes text parsing and recognition to enhance the model's perception

and structured understanding of textual content in images. Training involves approximately 50B tokens, with the learning rate warmed up from 3×10^{-4} to a peak value before decaying to 3×10^{-5}. We apply gradient clipping and mixed precision to ensure alignment stability across complex layouts and multilingual settings.

- **Stage-2:** In the second stage, all model parameters are unfrozen for end-to-end vision-language joint learning. The focus shifts to enhancing the model's capability for deep understanding and cognitive reasoning regarding structured content such as documents, tables, and charts. The training data mixture increases the proportion of synthetic samples covering multiple tasks—including text parsing, spotting, translation, and VQA—while retaining approximately ($\leq 10\%$) plain text to maintain instruction-following and linguistic generalization capabilities. This stage utilizes approximately 300B tokens with a warmup-cosine learning rate schedule, decaying from 2×10^{-4} to 5×10^{-5}. We stabilize convergence through progressive hard-example mixing, deduplication, and consistency filtering.
- **Stage-3:** In the third stage, we extend the model's context window to 32K by incorporating long-context parsing tasks and lengthy plain text data. This stage targets complex analytical scenarios, utilizing real-world auto-annotated data (spotting, translation, VQA), information extraction datasets, and long documents to model cross-page relationships and dense layouts. We further increase the share of structured output tasks (HTML/Markdown/LaTeX) and introduce a controlled fraction of difficult samples (e.g., text-free images, blur, distortions) to enhance robustness. This stage processes approximately 80B tokens, with the learning rate decaying from 8×10^{-5} to 5×10^{-6}. Stable training is achieved via gradient checkpointing, sequence parallelism, and efficient attention mechanisms.
- **Stage-4:** We conduct annealing training using carefully curated, human-annotated real-world data supplemented with a small proportion of high-quality synthetic samples, while maintaining the 32K context window. By employing unified instruction templates and standardized output formats across different tasks—such as spotting, document parsing, table parsing, formula recognition, information extraction, and translation—we ensure consistency in response patterns. This unified approach reduces the model's learning burden and facilitates the design of multi-task reward models in the subsequent post-training phase. The training utilizes 24B tokens, with the learning rate linearly decaying from 2×10^{-5} to 1×10^{-6}.

5.2 Reinforcement Learning (RL)

Reinforcement learning (RL) algorithms have emerged as a powerful paradigm, achieving remarkable success across various domains involving large language models (LLMs) and multimodal large language models (MLLMs). Notable applications include mathematical reasoning [Shao et al.(2024)Shao, Wang, Zhu, Xu, Song, Zhang, Li, Wu, and Guo], image segmentation [Liu et al.(2025)Liu, Peng, Zhong, Yue, Lu, Yu, and Jia], and omni-multimodal LLMs [Zhao et al.(2025)Zhao, Wei, and Bo]. This broad success is largely attributed to RL's ability to align model outputs with verifiable metrics [Wang et al.(2025b)Wang, Liu, Zheng, Xu, Ye, Wu, Liang, Wang, Li, Miao, et al.] or human preferences [Peng et al.(2025a)Peng, Wang, Tian, Yang, Wu, Xu, Zhang, Isobe, Hu, and Zhang, Peng et al.(2025b)Peng, Yang, Jiang, and Tian]. While RL has traditionally been applied to large-scale reasoning models, we investigate its application to lightweight OCR models that prioritize efficient and accurate text understanding. Leveraging the structured nature and inherent verifiability of many OCR tasks, we adopt Reinforcement Learning with Verifiable Rewards (RLVR) for closed-form

Table 2: Overview of the four-stage pre-training recipe for MahenOCR.

Stages	Stage-1	Stage-2	Stage-3	Stage-4
Purpose	Vision-Language Alignment	Multimodal Pre-traning	Long-context Pre-training	Application-oriented SFT
Trainable Parts	ViT & Adapter	All	All	All
Learning Rate	3e−4 → 3e−5	2e−4 → 5e−5	8e−5 → 5e−6	2e−5 → 1e−6
Training Tokens	50B	300B	80B	24B
Sequence Length	8k	8k	32k	32k
Data Composition	Pure Text, Synthetic Parsing and Recognition Data, General Image Caption Data	Pure Text, Synthetic Spotting, Parsing, Translation and VQA Data	Long Pure Text, Real-world Auto-annotated Data, Long Document Parsing Data, Information Extraction Data	Human-annotated Data, Hard-negative Data, Standardized Instruction Data.

tasks such as text spotting and document parsing. For more open-ended tasks like translation and text-centric VQA, we design reward mechanisms based on an LLM-as-a-judge approach. By integrating RLVR and LLM-as-a-judge techniques, we demonstrate that even lightweight models like MahenOCR can achieve significant performance improvements, opening new possibilities for edge and mobile applications.

5.2.1 Data Curation

Our data pipeline emphasizes **quality, diversity, and difficulty balance**. In terms of quality, we combine high-quality open-source and synthetic datasets, filtering them using LLM-based judging to ensure image–text alignment and removing tasks that are easily exploitable (e.g., simple multiple-choice). For diversity, we cover a broad range of OCR-related tasks and maintain sufficient exploration by discarding samples with low output diversity or zero reward variance. Finally, to balance task difficulty, we employ pass-rate filtering based on model samples, removing both trivial and unsolvable examples.

5.2.2 Reward Design

We adopt an **ability-adaptive reward design**, where each OCR-related task type has a tailored reward formulation that aligns with its output characteristics.

- **Spotting:** For text spotting tasks, which require joint text recognition and bounding box localization, the reward is computed as follows. Each predicted bounding box is first assigned to a ground-truth box by maximizing the Intersection over Union (IoU). The reward for each matched pair is then calculated as one minus the normalized edit distance between the predicted and ground-truth text strings. Any unmatched predictions or ground-truth boxes incur a penalty by contributing a reward of zero to the average. The final reward is the mean score across all evaluated pairs, providing a balanced measure of both localization and recognition accuracy.
- **Document Parsing:** Document Parsing aims to convert document images into structured formats containing textual content, mathematical formulas, and tables. The evaluation emphasizes both

structural integrity and content accuracy. The reward is computed based on the normalized edit distance between the model's output and the ground-truth reference.

- **VQA:** The reward is binary (1 or 0), based on whether the model's answer semantically matches the reference. The scoring model evaluates only content completeness and factual correctness, tolerating minor stylistic differences while enforcing strict alignment on key content elements.
- **Translation:** We use a soft reward scheme where a scoring LLM compares the generated output against the reference and assigns a score in the range [0, 5]. This raw score is then debias-normalized to [0, 1]. Crucially, this normalization is designed to expand the reward granularity in the mid-range (2–4), enabling the model to better capture subtle improvements and differences in translation quality.

5.2.3 Training Strategy

We adopt the Group Relative Policy Optimization (GRPO) algorithm as our main reinforcement learning framework. In each training iteration, GRPO samples a group of responses $(o_1, o_2, \cdots, o_G)$ for a given query (q) from the old policy ($\pi_{\theta_{\text{old}}}$) and updates the current policy (π_θ) by maximizing the objective:

$$\mathcal{L}_{GRPO}(\theta) = \mathbb{E}_{[q\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)]} \frac{1}{G}\sum_{i=1}^{G}\left[\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{\text{ref}}\right)\right] \quad (1)$$

where A_i represents the advantage computed from the group rewards, and $\mathbb{D}_{\text{KL}}$ is the KL-divergence term for regularization. The ϵ and β control clipping and the strength of KL penalties, respectively. To ensure stable and reliable training, we enforce length constraints and a strict format during reward computation. Specifically, any output that exceeds the maximum length is immediately assigned a reward of zero. Similarly, for structured tasks like spotting and document parsing, outputs that fail to follow the required schema are also directly penalized with zero reward. These constraints help the optimization process focus exclusively on valid, well-structured, and verifiable outputs, thereby guiding the model to learn accurate reasoning and formatting behavior under constrained conditions.

6 Evaluation

6.1 Spotting

To comprehensively evaluate the model's text spotting performance across diverse scenarios, we constructed a benchmark comprising nine categories: artistic text, document images, game screenshots, handwritten text, advertisement scenes, card/certificate/invoice images, screen captures, street view text, and video frames. Each category contains 100 images, forming a 900-image evaluation set. Based on this benchmark, we compared MahenOCR with traditional pipeline-based open-source models, leading commercial APIs, and general Vision-Language Models (VLMs). The results shown in Table 3 demonstrate that our approach achieves the best overall performance. Specifically, as an end-to-end VLM solution, MahenOCR significantly outperforms traditional pipeline-based methods. Furthermore, compared to general VLMs, our method achieves superior accuracy with substantially fewer parameters, demonstrating notable advantages in both computational efficiency and performance.

Table 3: Comprehensive evaluation of spotting ability on in-house benchmark.

Model Type	Model	Overall	Art	Doc	Game	Hand	Ads	Receipt	Screen	Scene	Video
Traditional methods	PaddleOCR	53.38	32.83	70.23	51.59	56.39	57.38	50.59	63.38	44.68	53.35
	BaiduOCR	61.90	38.5	**78.95**	59.24	59.06	66.70	**63.66**	68.18	55.53	67.38
General VLMs	Gemini-2.5-Pro	23.44	21.79	35.16	10.02	38.49	29.89	20.80	17.59	18.33	18.90
	Qwen3-VL-2B-Instruct	29.68	29.43	19.37	20.85	50.57	35.14	24.42	12.13	34.90	40.10
	Qwen3-VL-235B-A22B-Instruct	53.62	46.15	43.78	48.00	68.90	64.01	47.53	45.91	54.56	63.79
	Seed-1.6-Vision	59.23	45.36	55.04	59.68	67.46	65.99	55.68	59.85	53.66	70.33
OCR-Specific VLMs	**MahenOCR**	**70.92**	**56.76**	73.63	**73.54**	**77.10**	**75.34**	63.51	**76.58**	**64.56**	**77.31**

6.2 Parsing

We systematically evaluated the model's performance on document parsing using three benchmark datasets. First, we conducted experiments on OmniDocBench [Ouyang et al.(2024)Ouyang, Qu, Zhou, Zhu, et al.], a publicly available and comprehensive document parsing benchmark that includes a diverse set of digital and scanned documents covering formulas, tables, paragraphs, and various structural elements. Second, to further assess the model's robustness in real-world captured scenarios, we created a Wild version of OmniDocBench[2] by printing the original documents and re-capturing them under challenging conditions—such as manual folding, bending, and varying illumination—to simulate realistic distortions encountered in everyday document photography. Finally, we evaluated the model on DocML[3], our internally curated multilingual parsing dataset designed to assess robustness across multiple languages and acquisition settings. DocML spans both digital/scanned and real-world captured documents across 14 high-frequency non-Chinese/English languages, including German, Spanish, Turkish, Vietnamese, Korean, Malay, Portuguese, Russian, French, Indonesian, Thai, Italian, and Japanese. For both OmniDocBench and its Wild variant, we followed the official evaluation protocol described in [Ouyang et al.(2024)Ouyang, Qu, Zhou, Zhu, et al.] and report results for MahenOCR alongside other leading document parsing models. As shown in Table 4, MahenOCR achieves the highest overall performance on both the digital/scanned and real-world captured settings, demonstrating strong generalization across diverse document formats and acquisition conditions. Notably, despite its relatively compact 1B parameter size, MahenOCR outperforms larger specialized OCR or VLM-based parsing models. On DocML, we adopt an overall edit-distance–based score as the evaluation metric to comprehensively measure the accuracy and robustness of parsed outputs across multilingual settings. Under this metric, MahenOCR demonstrates excellent multilingual parsing performance, achieving state-of-the-art results across all 14 languages. These findings collectively show that MahenOCR delivers robust and accurate document parsing across multilingual, multi-scene, and real-world conditions.

6.3 IE & VQA

We systematically evaluate the model's performance on information extraction and open-ended visual question answering tasks using three benchmark datasets. First, to assess the model's capability on high-frequency card and document types, we constructed a test set comprising 768

[2]The Wild version of OmniDocBench will be publicly released in a future update.

[3]The DocML multilingual parsing dataset will also be open-sourced in a future release. We invite interested parties to reach out to us for access or evaluation prior to its public release.

Table 4: Parsing performance evaluated across multilingual settings and diverse document scenarios.

Model Type	Model	Size	OmniDocBench				Wild-OmniDocBench				DocML
			overall↑	text↓	formula↑	table↑	overall↑	text↓	formula↑	table↑	
General VLMs	Gemni-2.5-pro	-	88.03	0.075	85.92	85.71	80.59	0.118	75.03	78.56	82.64
	Qwen3-VL-235B	235B	89.15	0.069	88.14	86.21	79.69	0.09	80.67	68.31	81.40
Specialized VLMs (Modular)	MonkeyOCR-pro	3B	88.85	0.075	87.5	86.78	70.00	0.211	63.27	67.83	56.50
	MinerU2.5	1.2B	90.67	0.047	88.46	88.22	70.91	0.218	64.37	70.15	52.05
	PaddleOCR-VL	0.9B	92.86	0.035	91.22	90.89	72.19	0.232	65.54	74.24	57.42
Specialized VLMs (End2End)	Mistral-OCR	-	78.83	0.164	82.84	70.03	-	-	-	-	64.71
	Deepseek-OCR	3B	87.01	0.073	83.37	84.97	74.23	0.178	70.07	70.41	57.22
	dots.ocr	3B	88.41	0.048	83.22	86.78	78.01	0.121	74.23	71.89	77.50
	MahenOCR	1B	**94.10**	0.042	**94.73**	**91.81**	**85.21**	**0.081**	**82.09**	**81.64**	**91.03**

samples across 30 common categories (**??**), such as identification cards, passports, and invoices. Second, to evaluate text extraction performance in complex scenarios, we built a video subtitle dataset containing 1,000 samples covering diverse video contexts and subtitle styles. Additionally, the model was comprehensively evaluated on OCRBench [Liu et al.(2024a)Liu, Li, Huang, Yang, Yu, Li, Yin, Liu, Jin, and Bai], a publicly available benchmark that includes 1,000 test samples and spans multiple competencies, including scene text recognition, handwritten text and formula recognition, information extraction, and open-ended question answering on documents and charts. We evaluated the first two benchmarks using exact-match accuracy under a unified prompting protocol for multi-field JSON outputs, while adopting the official standard evaluation protocol for OCRBench. MahenOCR was compared against leading SOTA VLMs, including Qwen3VL-235B-Instruct, Seed1.6-VL-Instruct, and Gemini-2.5-Pro, using identical prompts and post-processing procedures such as JSON format parsing. As summarized in Table 5, MahenOCR achieves the highest overall accuracy across all 30 document categories in card/receipts information extraction and subtitle extraction tasks, despite having only around 1B parameters, significantly outperforming considerably larger VLMs such as Qwen3VL-235B-Instruct, Seed1.6-VL, and Gemini-2.5-Pro. On OCRBench, MahenOCR also demonstrates substantially better performance than DeepseekOCR at a similar scale and comparable with the larger Qwen3VL-2B-Instruct model. For OCRBench, we adopt the official standard evaluation protocol. We compare MahenOCR against mainstream SOTA VLMs such as Qwen3VL-235B-Instruct, Seed1.6-VL-Instruct, and Gemini-2.5-Pro using identical prompts and post-processing (JSON format parsing). As shown in Table 5, MahenOCR, despite having approximately 1B parameters, achieves the best overall performance across all 38 categories, substantially surpassing much larger VLMs such as Qwen3VL-235B-Instruct, Seed1.6-VL, and Gemini-2.5-Pro.

6.4 Text Image Translation

We systematically evaluated the model's text image translation capability using two benchmark datasets. For public benchmarking, we selected DoTA [Liang et al.(2024)Liang, Zhang, Ma, Zhang, Zhao, Xiang, Zong, and Zhou], a document translation dataset designed for complex and diverse layout scenarios, and used it to assess the model's end-to-end translation performance under realistic document conditions. In addition, we constructed an in-house evaluation benchmark based on DocML, where each sample is annotated with high-quality reference translations. This internal

Table 5: Evaluation of information extraction and visual question answering tasks.

Model	Cards	Receipts	Video Subtitles	OCRBench
DeepSeek-OCR [Wei et al.(2025)Wei, Sun, and Li]	10.04	40.54	5.41	430
PP-ChatOCR [PaddleOCR(2025)]	57.02	50.26	3.1	-
Qwen3-VL-2B-Instruct qwen3-vl	67.62	64.62	3.75	858
Seed-1.6-Vision [Seed(2025)]	70.12	67.5	60.45	881
Qwen3-VL-235B-A22B-Instruct qwen3-vl	75.59	78.4	50.74	**920**
Gemini-2.5-Pro comanici2025gemini	80.59	80.66	53.65	872
MahenOCR	**92.29**	**92.53**	**92.87**	860

benchmark enables a comprehensive assessment of translation robustness across multiple languages and a broad range of document types, including both digital/scanned and real-world captured scenes. To evaluate translation quality, we adopt the COMET [Rei et al.(2022)Rei, C. de Souza, Alves, Zerva, Farinha, Glushkova, Lavie, Coheur, and Martins] metric, a widely used neural-based evaluation standard for machine translation. As summarized in Table 6, MahenOCR surpasses VLMs with over 8B parameters on DoTA, demonstrating strong translation performance in complex document layouts despite its compact 1B scale. Furthermore, we achieved first place in the Track 2.2 OCR-free Small Model of the ICDAR 2025 Competition on End-to-End Document Image Machine Translation Towards Complex Layouts [Zhang et al.(2025)Zhang, Liang, Zhang, Chen, Xiang, Zhao, Zhou, and Zong], validating the effectiveness and generality of our approach. On the DocML evaluation set, MahenOCR again outperforms several larger VLMs exceeding 4B parameters, highlighting its robust multilingual translation capability across diverse layouts, languages, and acquisition conditions. These findings collectively demonstrate that MahenOCR provides a highly efficient yet powerful solution for text image translation in both public benchmarks and real-world multilingual scenarios. However, due to its relatively small language model, MahenOCR's translation capability lags behind its strong text detection, recognition, and document parsing performance. For applications requiring higher translation accuracy, developers can cascade our multilingual parsing module with larger translation models or await our upcoming general vision-language models to further boost overall translation quality.

7 Conclusion

In this report, we introduce MahenOCR, an open-source expert vision-language model that integrates a wide range of OCR tasks into a streamlined, end-to-end architecture. Our findings confirm that a compact 1B-parameter model can rival the performance of significantly larger general-purpose VLMs and conventional pipeline systems. This success validates the efficacy of our data-driven training methodology combined with targeted reinforcement learning strategies. MahenOCR secures state-of-the-art standing in key areas such as text spotting, document parsing, and information extraction, all while drastically reducing deployment complexity. These achievements fulfill our primary objective of harmonizing versatility with efficiency. Moving forward, we are committed to enhancing inference efficiency via advanced token compression and architectural refinements. We also plan to extend the model's capacity to process higher-resolution inputs and multi-page documents.

Table 6: Evaluation of photo translation. We additionally manually annotated DocML with high-quality English and Chinese reference translations to serve as ground-truth labels for evaluating text translation performance.

Model	**Size**	**DocML**		**DoTA**
		other2en	other2zh	en2zh
Gemni-2.5-Flash [Comanici et al.(2025)Comanici, Bieber, Schaekermann, Pasupat, Sachdeva, Dhillon, Blistein, Ram, Zhang, Rosen, et al.]	-	**79.26**	**80.06**	**85.60**
Qwen3-VL-235B-Instruct [Qwen(2025)]	235B	73.67	77.20	80.01
Qwen3-VL-8B-Instruct [Qwen(2025)]	8B	75.09	75.63	79.86
Qwen3-VL-4B-Instruct [Qwen(2025)]	4B	70.38	70.29	78.45
Qwen3-VL-2B-Instruct [Qwen(2025)]	2B	66.30	66.77	73.49
PP-DocTranslation	-	52.63	52.43	82.09
MahenOCR	1B	73.38	73.62	83.48

Our long-term vision focuses on adapting MahenOCR for deployment on edge devices, thereby democratizing access to robust OCR intelligence for a broader spectrum of real-world applications.

6.5pt plus 2pt

References

[Achiam et al.(2023)Achiam, Adler, Agarwal, Ahmad, Akkaya, Aleman, Almeida, Altenschmidt, Altman, Anadkat, et a Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Adeshola & Adepoju(2024)Adeshola and Adepoju] Ibrahim Adeshola and Adeola Praise Adepoju. The opportunities and challenges of chatgpt in education. *Interactive Learning Environments*, 32(10):6159–6172, 2024.

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

[Baidu(2025)] Baidu. BaiduOCRAPI. 2025. URL https://ai.baidu.com/tech/ocr/general.

[Blecher et al.(2023)Blecher, Cucurull, Scialom, and Stojnic] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

[Blecher et al.(2023)Blecher, Cucurull, Scialom, and Stojnic] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

[Brown et al.(2020)Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, et al.] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Comanici et al.(2025)Comanici, Bieber, Schaekermann, Pasupat, Sachdeva, Dhillon, Blistein, Ram, Zhang, Rosen, et a Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[Cortes & Vapnik(1995)Cortes and Vapnik] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[Cui et al.(2025a)Cui, Sun, Liang, Gao, Zhang, Liu, Wang, Zhou, Liu, Lin, et al.] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025a.

[Cui et al.(2025b)Cui, Sun, Lin, Gao, Zhang, Liu, Wang, Zhang, Zhou, Liu, et al.] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025b.

[Devlin et al.(2019)Devlin, Chang, Lee, and Toutanova] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. NAACL*, 1:4171–4186, 2019.

[Metanthropic(2025)] Metanthropic. Metanthropic-0.5B. 2025. URL https://github.com/Metanthropic/Metanthropic-0.5B.

[dots(2024)] dots. dots.ocr: Multilingual document layout parsing in a single vision-language model. 2024. URL https://github.com/rednote-hilab/dots.ocr.

[Eddy(1996)] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3): 361–365, 1996.

[Feng et al.(2025)Feng, Wei, Fei, Shi, Han, Liao, Lu, Wu, Liu, Lin, et al.] Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, et al. Dolphin: Document image parsing via heterogeneous anchor prompting. *arXiv preprint arXiv:2505.14059*, 2025.

[JaidedAI(2020)] JaidedAI. Easyocr. 2020. URL https://github.com/JaidedAI/EasyOCR.

[Khan & Umer(2024)Khan and Umer] Muhammad Salar Khan and Hamza Umer. Chatgpt in finance: Applications, challenges, and solutions. *Heliyon*, 10(2), 2024.

[Kuang et al.(2021)Kuang, Sun, Li, Yue, Lin, Chen, Wei, Zhu, Gao, Zhang, et al.] Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, et al. Mmocr: a comprehensive toolbox for text detection, recognition and understanding. *ACM Multimedia*, pp. 3791–3794, 2021.

[Li et al.(2021)Li, Qian, Yu, Qin, Zhang, Liu, Yao, Han, Liu, and Ding] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. *Proc. ACM Multimedia*, pp. 1912–1920, 2021.

[Li et al.(2023)Li, Lv, Chen, Cui, Lu, Florencio, Zhang, Li, and Wei] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. *Proc. AAAI*, 37(11):13094–13102, 2023.

[Li et al.(2025)Li, Liu, Liu, Ma, Zhang, Zhang, Guo, Zhang, Wang, and Bai] Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*, 2025.

[Liang et al.(2024)Liang, Zhang, Ma, Zhang, Zhao, Xiang, Zong, and Zhou] Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. Document image machine translation with dynamic multi-pre-trained models assembling. *Proc. NAACL*, pp. 7084–7095, 2024.

[Liao et al.(2017)Liao, Shi, Bai, Wang, and Liu] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. *Proc. AAAI*, 31(1), 2017.

[Liao et al.(2022)Liao, Zou, Wan, Yao, and Bai] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE TPAMI*, 45(1):919–931, 2022.

[Liu et al.(2023)Liu, Li, Wu, and Lee] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.

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

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

[Liu et al.(2024c)Liu, Feng, Xue, Wang, Wu, Lu, Zhao, Deng, Zhang, Ruan, et al.] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024c.

[Liu et al.(2025)Liu, Peng, Zhong, Yue, Lu, Yu, and Jia] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[Long et al.(2021)Long, He, and Yao] Shangbang Long, Xin He, and Cong Yao. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129(1):161–184, 2021.

[Lyu et al.(2018)Lyu, Liao, Yao, Wu, and Bai] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *Proc. ECCV*, pp. 67–83, 2018.

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

[Lyu et al.(2024b)Lyu, Zhang, Liu, Qiao, Xu, Wu, Yao, Han, Ding, and Wang] Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Maskocr: Scene text recognition with masked vision-language pre-training. *Transactions on Machine Learning Research*, 2024b.

[Metanthropic(2025)] Metanthropic. Metanthropic-0.5B. 2025. URL https://github.com/Metanthropic/Metanthropic-0.5B.

[Niu et al.(2025)Niu, Liu, Gu, Wang, Ouyang, Zhao, Chu, He, Wu, Zhang, et al.] Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, et al. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*, 2025.

[Ouyang et al.(2024)Ouyang, Qu, Zhou, Zhu, et al.] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations, 2024. URL https://arxiv.org/abs/2412.07626.

[PaddleOCR(2025)] PaddleOCR. Pp-chatocr. 2025. URL https://github.com/PaddlePaddle/PaddleOCR.

[Peng et al.(2025a)Peng, Wang, Tian, Yang, Wu, Xu, Zhang, Isobe, Hu, and Zhang] Shangpin Peng, Weinong Wang, Zhuotao Tian, Senqiao Yang, Xing Wu, Haotian Xu, Chengquan Zhang, Takashi Isobe, Baotian Hu, and Min Zhang. Omni-dpo: A dual-perspective paradigm for dynamic preference learning of llms. *arXiv preprint arXiv:2506.10054*, 2025a.

[Peng et al.(2025b)Peng, Yang, Jiang, and Tian] Shangpin Peng, Senqiao Yang, Li Jiang, and Zhuotao Tian. Mitigating object hallucinations via sentence-level early intervention. *arXiv preprint arXiv:2507.12455*, 2025b.

[Qin et al.(2024)Qin, Huang, Wang, Zhang, Wang, Liang, Wang, Lin, He, and Zhang] Qintong Qin, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang, Shawn Wang, Matthieu Lin, Conghui He, and Wentao Zhang. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *CoRR*, abs/2410.21169, 2024.

[Qwen(2025)] Qwen. Qwen3-vl. 2025. URL https://github.com/QwenLM/Qwen3-VL.

[Radford et al.(2019)Radford, Wu, Child, Luan, Amodei, Sutskever, et al.] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Rei et al.(2022)Rei, C. de Souza, Alves, Zerva, Farinha, Glushkova, Lavie, Coheur, and Martins] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proc. WMT*, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52/.

[Seed(2025)] Seed. Seed1.6. 2025. URL https://seed.bytedance.com/en/seed1_6.

[Shao et al.(2024)Shao, Wang, Zhu, Xu, Song, Zhang, Li, Wu, and Guo] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

[Shi et al.(2017)Shi, Bai, and Yao] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2017.

[Shi et al.(2018)Shi, Yang, Wang, Lyu, Yao, and Bai] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE TPAMI*, 41(9):2035–2048, 2018.

[Su et al.(2024)Su, Ahmed, Lu, Pan, Bo, and Liu] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[Sun et al.(2025)Sun, Cui, Du, and Liu] Ting Sun, Cheng Cui, Yuning Du, and Yi Liu. Pp-doclayout: A unified document layout detection model to accelerate large-scale data construction. *arXiv preprint arXiv:2503.17213*, 2025.

[Team(2025)] Qwen Team. Qwen3 technical report. 2025. URL https://arxiv.org/abs/2505.09388.

[Tschannen et al.(2025)Tschannen, Gritsenko, Wang, Naeem, Alabdulmohsin, Parthasarathy, Evans et al.] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[Wang et al.(2025a)Wang, Gao, Gu, Pu, Cui, Wei, Liu, Jing, Ye, Shao, et al.] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

[Wang et al.(2025b)Wang, Liu, Zheng, Xu, Ye, Wu, Liang, Wang, Li, Miao, et al.] Xumeng Wang, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025b.

[Wei et al.(2025)Wei, Sun, and Li] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

[Yin et al.(2024)Yin, Fu, Zhao, Li, Sun, Xu, and Chen] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[Yu et al.()Yu, Li, Zhang, Zhang, Guo, Qin, Yao, Han, Ding, and Wang] Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Structextv2: Masked visual-textual prediction for document image pre-training. *The Eleventh International Conference on Learning Representations*.

[Zhang et al.(2024)Zhang, Wang, Huang, Zhang, Wang, Liang, He, and Zhang] Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*, 2024.

[Zhang et al.(2025)Zhang, Liang, Zhang, Chen, Xiang, Zhao, Zhou, and Zong] Yaping Zhang, Yupu Liang, Zhiyang Zhang, Zhiyuan Chen, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. Icdar 2025 competition on end-to-end document image machine translation towards complex layouts. In *International Conference on Document Analysis and Recognition*, pp. 505–522. Springer, 2025.

[Zhao et al.(2024)Zhao, Kang, Wang, and He] Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*, 2024.

[Zhao et al.(2025)Zhao, Wei, and Bo] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

[Zhou et al.(2017)Zhou, Yao, Wen, Wang, Zhou, He, and Liang] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: an efficient and accurate scene text detector. In *CVPR*, pp. 2642–2651. IEEE Computer Society, 2017.

[Unknown(2025)] Mistral OCR: Free online ai ocr tool to extract text. https://www.mistralocr.com/, 2025. Accessed: 2025-07-30.

[Wang et al.(2025)Wang, Hu, Li, Safari, and Yang] Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*, 2025.