PP-DocBee: Improving Multimodal Document Understanding Through a Bag of Tricks

Feng Ni† **Kui Huang**† **Yao Lu** **Wenyu Lv** **Guanzhong Wang** **Zeyu Chen** **Yi Liu**
PaddlePaddle Team, Baidu Inc.
nifeng03@baidu.com huangkui01@baidu.com lvwenyu01@baidu.com liuyi22@baidu.com

Abstract

With the rapid advancement of digitalization, various document images are being applied more extensively in production and daily life, and there is an increasingly urgent need for fast and accurate parsing of the content in document images. Therefore, this report presents PP-DocBee, a novel multimodal large language model designed for end-to-end document image understanding. First, we develop a data synthesis strategy tailored to document scenarios in which we build a diverse dataset to improve the model generalization. Then, we apply a few training techniques, including dynamic proportional sampling, data preprocessing, and OCR postprocessing strategies. Extensive evaluations demonstrate the superior performance of PP-DocBee, achieving state-of-the-art results on English document understanding benchmarks and even outperforming existing open source and commercial models in Chinese document understanding. The source code and pre-trained models are publicly available at https://github.com/PaddlePaddle/PaddleMIX.

1 Introduction

Recent advances in multimodal large language models (MLLMs) (Liu et al., 2023e;c; Zhu et al., 2023; Bai et al., 2023b) have demonstrated remarkable capabilities in general vision-language understanding through alignment of visual encoders (Dosovitskiy et al., 2021; Radford et al., 2021) with Large Language Models (LLMs) (Touvron et al., 2023; Vicuna, 2023; Bai et al., 2023a). Specifically, ViT is tasked with image processing to extract visual features. These features are subsequently processed and integrated by the MLP. The LLM component is responsible for understanding and generating text. This synergistic combination enables the model to process image and text information concurrently, facilitating the understanding of comprehensive multimodal documents.

Popular open source MLLM models such as Qwen2-VL (Wang et al., 2024) and InternVL2 (Team, 2024) follow the "ViT+MLP+LLM" paradigm and score high on the leaderboards of some document understanding tasks. However, these models exhibit significant limitations when processing text-rich visual content such as documents, tables, and charts (Liu et al., 2023g), mainly due to two inherent constraints: (1) existing vision-to-text modules are optimized for natural image features rather than textual/structural representations, and (2) the current document understanding model has weak ability in Chinese scenarios.

Document image understanding has excellent potential in various fields. For example, corporate workflows require automatic extraction of financial data and contract terms, academic research requires efficient analysis of various articles and archives, and individuals must also process some receipts/forms intelligently. Although traditional approaches (Li et al., 2022) that combine optical character recognition (OCR) with rule-based systems have partially succeeded, they struggle with complex document layouts, poor image quality, and contextual reasoning requirements.

To effectively address these challenges, we introduce PP-DocBee, with the two main contributions elucidated below:

†Equal Contribution.

- We propose a document data synthesis strategy that generates a 477k high-quality dataset on Chinese document understanding named PPInfinityDocData. We also collected a public dataset of 3.3M from various document sources.
- We validate the effectiveness of the data strategy on an open-source multimodal large model. The experiments demonstrate that the generic model can improve document understanding through a few techniques with our data. It achieves state-of-the-art performance, surpasses its counterparts, and maintains a lightweight and fast architecture.

2 Data Synthesis Strategy

In terms of data quality, existing open-source document datasets have significant deficiencies. The lack of Chinese corpora, uneven quality of images and texts, absence of information extraction capabilities, and insufficient scene diversity are particularly prominent. Existing multimodal large models face multiple limitations while generating question-answer pairs regarding data generation. These limitations include systemic defects such as uncontrollable generation costs, poor answer accuracy, and deviation from the focus of the question. These problems directly contribute to the inability to understand multimodal large language models (MLLMs) for Chinese documents.

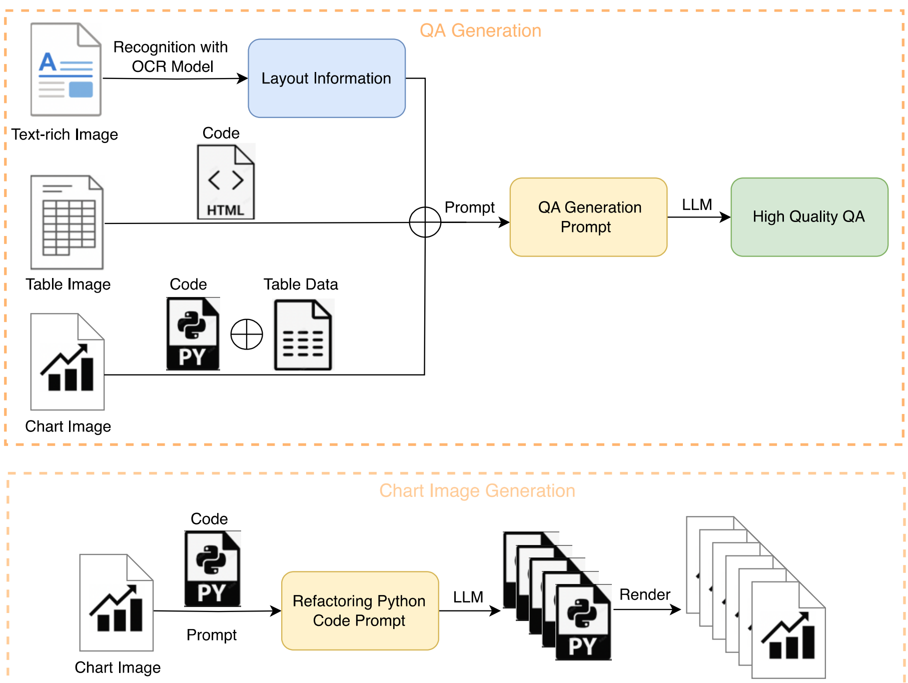


Figure 1: Document image QA generation and Chart image generation pipeline.

We propose a data generation pipeline for document understanding to improve the Chinese semantic understanding and expand the coverage of multimodal scenarios, as shown in Figure 1. This pipeline targets three typical document types: text-rich documents, tables, and charts, and designs differentiated data generation pipelines. A multimodal collaborative generation mechanism balances data quality and generation efficiency. The core technical strategies include: (1) constructing a cascaded processing architecture that combines a small Optical Character Recognition (OCR) model and a large language model, which can collaborate to make up for shortcomings; (2) developing a rendering engine based on semantic control, which helps expand Chinese chart images; and (3) designing document type-sensitive prompt engineering templates, which helps generate high-quality Chinese document question-answer(QA) pairs.

2.1 TEXT-RICH DOCUMENT

Text-rich documents are a common document type that contains a large amount of text and a small amount of pictures, tables, and charts. This data type helps the model understand the overall semantics of the document and general questions.

We compared the strategy of generating QA based on MLLMs directly from images with the strategy of generating QA based on an LLM in combination with a smaller OCR model. The strategy that uses MLLMs for generating QA can lead to knowledge expansion when generating instructions, which involves issues like generating instructions based on non-document knowledge. Additionally, this strategy often focuses on visually perceptible information and unconventional OCR-related questions. We provide an example of the text-rich document image and QA pairs generation shown in Figure 2.

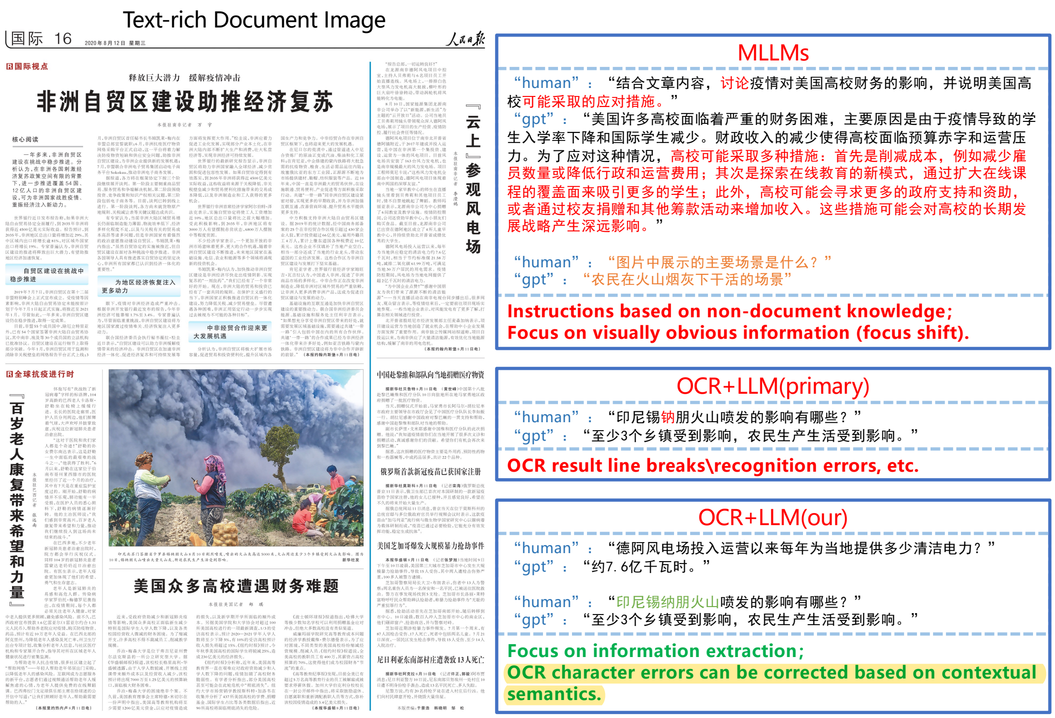


Figure 2: Text-rich document QA Generation. Red and orange indicate errors and inappropriate generation, and green indicates generation with fixed issues. The image comes from the open source dataset M6Doc Cheng et al. (2023).

Text-rich Images. We filter complex-layout Chinese PDF documents from professional fields such as academic papers, financial reports, and research reports. Using document parsing tools, we construct a single-page document image dataset that includes mixed elements such as text, images, and formulas, preserving the layout features of the original documents.

QA Generation. We propose an OCR-LLM collaborative verification mechanism, which utilizes the PaddleOCR Li et al. (2022) to extract accurate layout structures and textual information. Subsequently, the OCR outputs are integrated with the semantic understanding capabilities of a Large Language Model (ERNIE-Bot 4.0): 1) parse document images and obtain the corresponding layout information '{json_string}' by using PaddleOCR, and the images are converted into text information. 2) design appropriate prompts to enable the LLM to correct OCR recognition errors based on the context semantics; 3) control the distribution of generated question-answer pairs through instruction templates. The prompt used is as follows:

Prompt for Text-rich Document QA Generation

question = f"You are a document data visual question-answering dialogue generation system. Your main task is to design instructions and corresponding answers based on the OCR layout information of the provided document image, so that when the obtained multimodal data is used for model training, the model can fully learn the multimodal document understanding ability.
Here is the OCR layout information extracted from a research report document image: {json_string}. Please imagine that you are looking at the corresponding image instead of a simple json string based on the content (text-related characters are recognized by line, please fully splice and understand it), and then combine the information in the image to design instructions and answers from the perspective of professional research report readers.
When generating these dialogues, please make sure to follow the following guidelines:
These instructions must meet the following requirements:
1. Instructions focus on the ability to extract information from document images, and the answers can be directly observed from the image.
2. Instructions and answers must be highly relevant to the image, and instructions cannot be answered without the image. If the instruction provides too much information from the image, so that the instruction can be answered without the image, it is strictly prohibited.
3. The instructions must be generated based on the information in the image, and the accurate answer can be obtained in combination with the image content.
4. The text characters in the provided OCR layout information contain the exact answer to the answer. Please strictly ensure that the answer is correct, otherwise do not generate the instruction and answer.
5. For each layout area type (printed text, tables, charts, printed formulas, seals), if the document contains this type of information, then please generate instructions based on this type of content, otherwise no need to generate.
6. The answer to the instruction should be as concise and accurate as possible. Do not repeat the question and reply directly to the answer. At the same time, ensure that the answer is directly obtained from the original text of the image, and do not summarize.
7. Instructions and questions should not contain information about the layout structure.
8. Instructions should directly give questions, and do not use words such as 'Please ask', 'Please answer', and 'In the document'.
9. When generating instructions related to tables, if the table contains relevant information such as units, percentages, positive and negative signs, please ensure the completeness of the answer.
Please generate at least 5 Chinese instructions and answers. You need to provide the generated content in JSON format. Please make sure that ```json``` is included in the output. You can refer to the following sample to organize your output: {template}. "

2.2 CHART

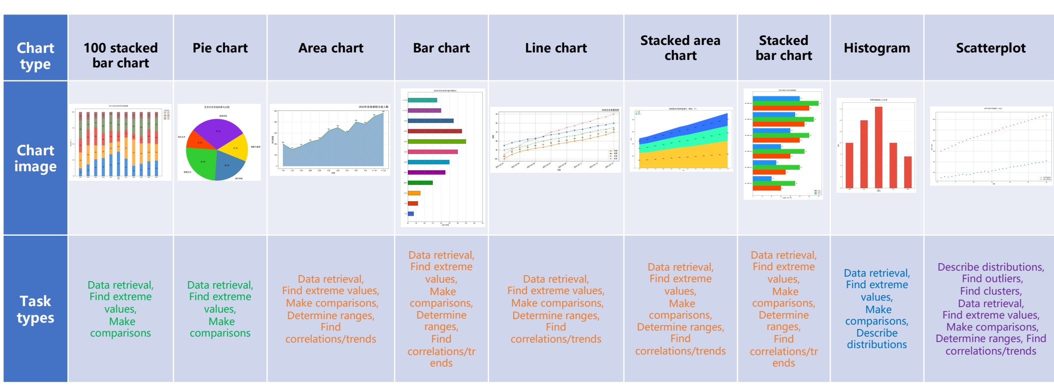


Figure 3: Common chart types and the corresponding task types

For chart data, MMLMs generally perform poorly in analyzing charts, and there are fewer high-quality Chinese chart images. So we summarize the currently common nine types of chart types as Figure 3, and design an image generation and QA generation scheme for each type of chart, and summarize the corresponding task types that can be asked.

Chart Images. We obtain 2,159 high-quality chart data(including images, codes and data tables) as our seed data by manual screening and filtering out incomplete image information and data with poor text-image relevance from the English chart dataset open-sourced by LLaVA-Chart Li et al. (2022), and develop a parametric Chinese chart image generation engine: 1) use LLM to semantically modify parameters in code corresponding to the chart (e.g., values, axes, colors, legends, themes), and randomly select and change the theme of the chart from the topics pool (e.g., "Art & Design", "Science & Nature"); 2) translate the text presented on the image into Chinese, when the LLM outputs code, while keeping other parameters unchanged. 3) prevent rendering issues such as image annotations exceeding the image range when the generated code is rendered into an image by fixing the code format; 4) parse the generated text into Python code and then render it into an image using Matplotlib. In this way, we obtain many Chinese chart images with rich scenarios. The prompt used is as follows, area chart for example:

Prompt for Chart Image Generation (area chart)

question = f"You are a highly intelligent AI familiar with data visualization and matplotlib. Given the following matplotlib code of an area chart: {code}. Please generate a diversified version of this matplotlib code of area chart. Here are some options you should follow:
1. Generate data points that fit the chart and the number of data points should be as much as you can to provide, but you should not skip any data point in your code. Do not add any comment in your code.
2. Add or change some data point with new corresponding values to enrich the visualization and print the final table data you use in the code within triple backticks (```), and don't include things like csv, plaintext in the triple backticks (```).
3. Modify the color scheme using specific color codes (e.g., #RRGGBB) for better clarity or visual appeal. Avoid using color categories.
4. Change width and height of the chart reasonably.
5. Change the topic, headline, and data type (which fit the topic) of the chart, put the headline in an appropriate place that does not overlap with other things (make sure the headline does not overlap with the legend! Put these two things away), you can refer and choose one (not all) of the topics from: {topics_pool}, but reduce using global topic, do not use temperature.
6. Generate only an area chart with one column of value, not a stacked area chart.
7. Assign annotation/text label on the chart. Do not use random data.
You should choose some of these options, not all of them, to diversify the visualization.
Different data points should have different values.
You should give FULL code with ALL data points and don't miss any detail.
Make sure the legend appears completely in the chart after the code is rendered.
Print table data first in Chinese, the table data format should be able to directly write in csv file, then print your code (keep the topic, headline, and data type (which fit the topic) of the chart in Chinese). Include the code with ``` ``` format."

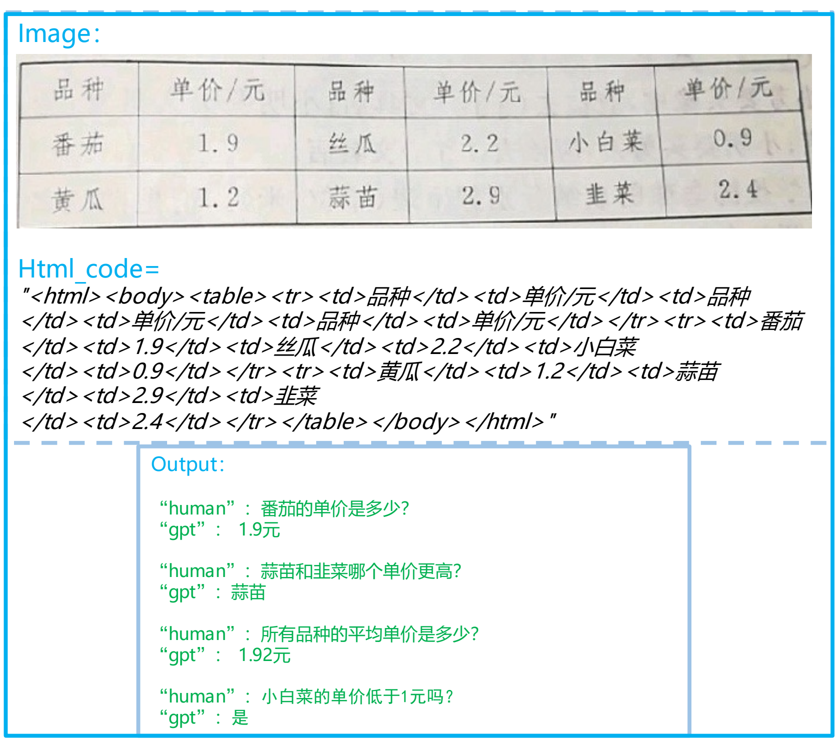
QA Generation. With the help of the code and data tables corresponding to the chart image, QA generation will become more accurate. We design a data-chart driven question-answer generation framework: 1) extract statistical features from the code and data tables as question material; 2) match preset question templates based on chart types and task types as Figure 3; 3) use LLM for semantic expansion and logical verification to generate QA pairs in Chinese that include professional questions and exact answers with the format as the template. The prompt used is as follows:

Prompt for Chart QA Generation

question = f"You are a highly intelligent AI familiar with data visualization and {chart_type}.
Below is the matplotlib code for the {chart_type} chart: {code} and the corresponding table data: {table_data}.
Please imagine that you are looking at the image generated by the code, not the code itself.
Please generate questions of different task types based on the content of the chart.
The task type is {task_types}.
Remember that in your answer, only the image of the chart is given, and your answer is based on the image. The table data is the real value of the relevant numerical value in the image, so make sure the answer is correct.
The value and label of the question are the real basis of your question, so make sure the answer is correct.
Avoid using invalid escape characters in strings.
Use approximate color names instead of hexadecimal colors. If there are units, % and other information in the chart, please ensure the integrity of the units , % and other information in the answer.
In addition, I hope to save your output as a json file, so I hope you can organize your answer like {template}, and make sure the values of human and gpt in json are all in Chinese. "

2.3 Table

In addition to chart data, table data is important in document understanding. Similar to charts, high-quality Chinese table data suitable for multimodal large model training is also lacking. Therefore, we have designed a high-quality QA production strategy based on the internal table data used for layout analysis tasks (including table data and HTML code). We do not produce many separate data because the text-rich data will contain some table data.



品种	单价/元	品种	单价/元	品种	单价/元
番茄	1.9	丝瓜	2.2	小白菜	0.9
黄瓜	1.2	蒜苗	2.9	韭菜	2.4

Figure 4: Example for table QA generation

Table Images. Based on the internal table data used for layout analysis tasks (including table data and HTML code), we construct a dataset.

QA Generation. We establish a HTML-Table dual-modal alignment mechanism, using the original HTML table structure as baseline information to design hierarchical prompt templates: 1) extract table themes and statistical features for html code, which ensures that the model understands and grasps the values in the table; 2) generate questions that conform to cognitive logic with pre-set task types, which prevents outputting QA that do not conform to human preferences. An example for table QA generation is as Figure 4. The prompt used is as follows:

Prompt for Table QA Generation

question = f"This is a table chart displayed by html code: {html_code}. Please generate questions of different task types about the chart, and make sure that the answers to the questions can be clearly obtained from the chart, otherwise you can choose not to select the corresponding task.
The selectable task types and their explanations are as follows:
Factoid: Asks a specific fact, the answer comes from a fragment in the table or a value obtained by aggregation.
Free Form: The answer has no fixed format, is usually longer, and is similar to a conversation, such as ChatGPT.
Multiple Choice: The question requires selecting one of multiple options as an answer.
List: The answer requires providing a series of related items, usually from multiple rows or columns in a table.
Yes/No: The answer to the question can only be 'yes' or 'no'.
Explanation: The answer requires explaining specific data or trends in the table.
Comparison: The answer requires comparing two or more values in the table.
Causal: The answer requires explaining the causal relationship between the data in the table.
Computation: The answer requires some calculations, such as sum, average, etc.
Classification: According to the question, classify or categorize some data in the table.
Time Series: The answer requires analyzing the time series data in the table, such as trends, patterns, etc.
You need to imagine that you are looking at an image rendered by the code, not the code itself. Remember in your answer, all you are given is an image of a graph, and you are answering based on the image.
The values and labels are the ground truth of your question, so make sure the answer is correct. Avoid using invalid escape characters in strings. Also, I want to save your output to a json file, so I want you to organize your answer like {template} and must include the answer in ```json``` format."

2.3.1 PPINFINITYDOCDATA

Owing to the intelligent data generation pipeline, we have successfully generated and screened 477k high-quality Chinese multimodal data, which are collectively named PPInfinityDocData. This dataset has made a significant contribution to the improvement of the final model training. The overall distribution of the synthetic dataset is illustrated in Figure 5.

PPInfinityDocData has three categories: Text-rich Document(Doc), Chart and Table. The specific distribution of each category and its subcategories is shown in the figure: the number of Doc is 288k, accounting for 60% of the total dataset, the number of table is 26k, and the number of Chart is 163k. All Chinese data is 314k, accounting for 66% of the total dataset.

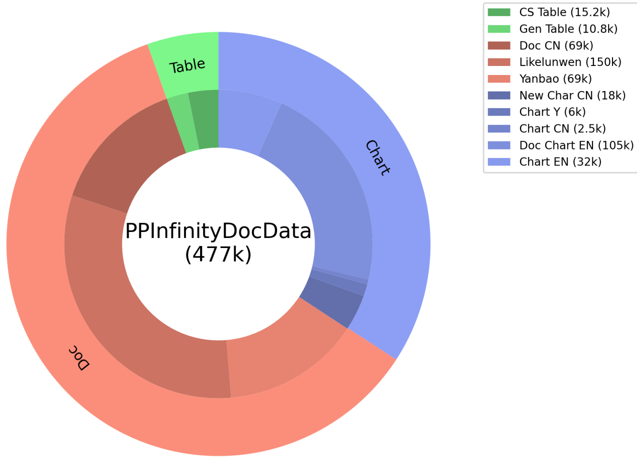


Figure 5: The overall distribution of Synthetic Dataset

3 Methodology

3.1 Model Overview

We chose Qwen2-VL (Wang et al., 2024; Authors, 2023) as the basic model because of its higher comprehensive accuracy. To deploy quickly and facilitate more low-resource users, we chose the Qwen2-VL-2B model to build our PP-DocBee. Although Qwen2-VL-2B has strong document comprehension capabilities, its proficiency is largely confined to English scenarios, given the predominance of English-language training data. In Chinese scenarios, its performance is significantly less developed, indicating substantial potential for improvement.

Applying our ”Bag-of-Freebies” technique did not alter the number of parameters or the structure of the basic model. Despite these constraints, it significantly improved the model capability to process English documents and a remarkable enhancement in its ability to handle Chinese documents.

3.2 Data Pre-processing

When processing an image, Qwen2-VL employs a patch-based approach, dividing the image into multiple small patches, each typically sized 28×28 pixels, similar to the Vision Transformer (ViT). This method decomposes the image into a series of visual tokens, enabling the model to better understand and process the image content. Notably, the length of Qwen2-VL’s visual tokens is dynamic, and the model dynamically adjusts the number of visual tokens based on the resolution of the input image. This adaptive mechanism is particularly effective for Visual Question Answering (VQA) tasks, as it provides more sufficient and comprehensive visual features.

During the training phase, we implemented an expanded range of resize thresholds, increasing the upper limit from 512 pixels to 768 pixels. This approach was designed to augment the overall resolution distribution of the datasets, thereby enriching the visual feature spectrum available to the model. In the inference stage, conventional-resolution images were proportionally upscaled by a factor ranging from 1.1 to 1.3. In contrast, images with lower resolutions maintained their original pre-processing strategies. These strategies provided a more comprehensive set of visual features, thereby enhancing the model’s understanding capabilities.

3.3 Dynamic Ratio Sampling Training

Our training data encompasses diverse document understanding datasets, including general VQA images, OCR images, charts, rich-text documents, mathematical and complex reasoning tasks, synthetic data, and plain text data. We implemented a dynamic data ratio sampling mechanism to optimise the training process and assign different sampling weights to different data and sources. This

Category	Data Composition
Mathematics	GeoQA+ Cao & Xiao (2022), MathQA Yu et al. (2023a), CLEVR-Math/Super Lindström & Abraham (2022); Li et al. (2023), Geometry3K Lu et al. (2021a), MAVIS-math-rule-geo Zhang et al. (2024), MAVIS-math-metagen Zhang et al. (2024), GEOS Seo et al. (2015), UniGeo Chen et al. (2022)
Science	AI2D Kembhavi et al. (2016), ScienceQA Lu et al. (2022a), TQA Kembhavi et al. (2017), VisualWebInstruct TIGER-Lab (2024)
Chart & Table	ChartQA Masry et al. (2022a), MMC-Inst Liu et al. (2023b), DVQA Kafle et al. (2018), PlotQA Methani et al. (2020), LRV-Instruction Liu et al. (2023a), TabMWP Lu et al. (2022b), UniChart Masry et al. (2023), TAT-DQA Zhu et al. (2022), FigureQA Kahou et al. (2017), Chart2Text Kantharaj et al. (2022), RobuT-{Wikisql, SQA, WTQ} Zhao et al. (2023)
Naive OCR	SynthDoG Kim et al. (2022), MTWI He et al. (2018), LVST Sun et al. (2019), FUNSD Jaume et al. (2019), Latex-Formula OleehyO (2024), IAM Marti & Bunke (2002), Handwriting-Latex aidapearson (2023), ArT Chng et al. (2019), CTW Yuan et al. (2019), ReCTs Zhang et al. (2019), COCO-Text Veit et al. (2016), SVRD Yu et al. (2023b), MapText Li et al. (2024b), CAPTCHA parasam (2024), Est-VQA Wang et al. (2020), HME-100K TAL (2023), TAL-OCR-ENG TAL (2023), TAL-HW-MATH TAL (2023), IMGUR5K Krishnan et al. (2023), Invoices-and-Receipts-OCR mychen76 (2024), IIIT5k Mishra et al. (2012), K12-Printing TAL (2023), Handwritten-Mathematical-Expression Azu (2023), WordArt Xie et al. (2022), Handwriting-Forms ift (2024)
OCR QA	DocVQA Clark & Gardner (2018), InfoVQA Mathew et al. (2022b), TextVQA Singh et al. (2019a), ArxivQA Li et al. (2024a), ScreenQA Hsiao et al. (2022), DocReason Hu et al. (2024), Ureader Ye et al. (2023), FinanceQA Sujet AI (2024), DocMatrix Laurençon et al. (2024), A-OKVQA Schwenk et al. (2022), Diagram-Image-To-Text Kamizuru00 (2024), MapQA Chang et al. (2022), OCRVQA Mishra et al. (2019), ST-VQA Biten et al. (2019), SQuAD-VQA, VQA-CD Mahamoud et al. (2024), MTVQA Tang et al. (2024)
General VQA	LLaVA-150K Liu et al. (2023d), LVIS-Instruct4V Wang et al. (2023), ALLaVA Chen et al. (2024a), Laion-GPT4V LAION (2023), LLAVAR Zhang et al. (2023), VizWiz Gurari et al. (2018), MMinstruct Liu et al. (2024a), WildVision Lu et al. (2024), LLaVA-Critic-113k Xiong et al. (2024), VQAv2 Goyal et al. (2017), MMRA Wu et al. (2024), MMDU Liu et al. (2024b), IconQA Lu et al. (2021b)
Text-only	WizardLM Xu et al. (2023), Infinity-Instruct BAAI (2024), UltraInteract-sft Yuan et al. (2024)

(a) Summary of the open source public dataset used in PP-DocBee.

Category	Total number of samples	Data Composition
Text-rich Document	288k	Financial Reports, Research Reports, Laws and Regulations, Science and Engineering Papers
Table	26k	Instructions, Science and Engineering Papers
Chart	163k	Contracts, Papers

(b) Summary of the synthetic dataset PPInfinityDocData used in PP-DocBee.

Table 1: SFT Dataset used in PP-DocBee.

approach significantly improves the training ratio of high-quality data and balances the quantitative differences between different datasets.

3.4 OCR Post-process

Optical Character Recognition (OCR) tools or models are employed to pre-extract text from images through OCR recognition. The extracted text is subsequently provided as auxiliary prior information for the image question. Specifically, the OCR-recognized text is incorporated into the input of the PP-DocBee model during the inference stage by adding a prompt to the original question: ”Use the image and the OCR result as context and answer the following question: ”. This method effectively enhances model performance, particularly on images containing clear and limited text.

4 Experiments

4.1 Training Data

The SFT datasets used in PP-DocBee are detailed in Table 1. This comprehensive collection includes a diverse array of public datasets, as described in Table 1a, as well as internally generated synthetic datasets, which are described in Table 1b. Integrating these datasets ensures a rich and varied training environment and improves the generalization ability across different scenarios and modalities.

4.2 Implementation Details

PP-DocBee is initialized from the Qwen2-VL-2B model (Wang et al., 2024), which employs a Vision Transformer (ViT) with approximately 675 million parameters as the visual encoder and a 1.5B Qwen2 Large Language Model (LLM) as the language decoder. During the supervised fine-tuning

Method	DocVQA-test	ChartQA	InfoVQA-test	TextVQA	OCRBench
Closed-source MLMMs					
GPT-4o OpenAI (2024)	92.8	85.7	79.2	77.4	73.6
Gemini-1.5-Pro Team et al. (2023)	93.1	87.2	80.1	78.7	75.4
Open-source MLMMs					
MiniCPM-V-2-2B Yao et al. (2024)	71.9	-	-	74.1	60.5
Aquila-VL-2B Gu et al. (2024)	85.0	76.5	58.3	76.4	77.2
Mini-Monkey-2B Huang et al. (2024)	87.4	76.5	60.1	76.0	79.4
InternVL2-2B Chen et al. (2024c)	86.9	76.2	58.9	73.4	78.1
InternVL2.5-2B Chen et al. (2024b)	88.7	**79.2**	60.9	74.3	80.4
Qwen2-VL-2B Wang et al. (2024)	90.1	73.5	65.5	79.7	80.9(82.2)
PP-DocBee-2B	**90.6**	74.6	**66.2**	**81.2**	**82.8(83.5)**

Table 2: Evaluation of existing OCR-Free MLLMs on public benchmarks. The 83.5 in brackets after 82.8 means the score after using OCR post-processing assistance.

(SFT) stage, the visual encoder is frozen while the parameters of the LLM are updated. PP-DocBee is trained for 16k iterations on a dataset comprising nearly 5 million samples, with a batch size of 32. This training process takes approximately 2 days using single node with 8 NVIDIA A800 GPUs.

4.3 Comparison with SOTA

We evaluated the performance of PP-DocBee on five English text-rich image benchmarks and our internal Chinese business scenario image benchmarks. PP-DocBee was compared with existing state-of-the-art OCR-free multimodal large language models (MLLMs) of the same parameter size, as well as closed-source APIs. The five English benchmarks cover documents (DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022a)), charts (ChartQA (Masry et al., 2022b)), natural images (TextVQA (Singh et al., 2019b)), and OCR-related images (OCRBench (Liu et al., 2023f)). Our internal Chinese business evaluation set includes financial reports, laws and regulations, scientific and engineering papers, instructions, liberal arts papers, contracts, research reports, and other relevant scenes. The resolution of all images is very high, with an average resolution of approximately 1680×1204, comprising a total of 1196 data samples. These can be categorized into four types: printed text (656 images), tables (358 images), seals (15 images), and charts (167 images).

As shown in Table 2, We compare PP-DocBee with the existing state-of-the-art OCR-free MLLMs and APIs. The table lists the performance indicators of multiple open-source and closed-source models on five tasks: DocVQA, ChantQA, InfoVQA, TextVQA, and OCRBench. By comparing these indicators, we found that the PP-DocBee-2B model showed excellent performance on multiple tasks, especially on the TextVQA task, which achieved a high score of 81.2, and on the OCRBench task, it also achieved a high score of 82.8; both of which are the highest scores among all models in the table. After using OCR post-processing assistance in the OCRBench task, we got a higher score of 83.5. This strategy also proved effective on Qwen2-VL, as shown in the table, from 80.9 points to 82.2 points. We found this strategy effectively enhances images containing clear and limited text. Considering the performance of all tasks, the PP-DocBee-2B model performed the best with comprehensive accuracy, which shows that it has high accuracy and reliability when dealing with Chinese-related multimodal tasks.

Table 3 illustrates the performance of the PP-DocBee-2B model across multiple categories. Notably, in the"Painted text" category, the PP-DocBee-2B model achieved a leading score of 517, outperforming other models. Furthermore, the PP-DocBee-2B model achieved a high score of 202 in the "Tables" category, demonstrating its proficiency in understanding and processing tabular data. Although the scores in the"Seals" and"Charts" categories were slightly lower, at 5 and 41, respectively, the overall score of 765 was the highest among all models. This result indicates that PP-DocBee-2B exhibits a high comprehensive accuracy when processing Chinese multimodal data, and it has demonstrated strong capabilities in text recognition, table parsing, and understanding other visual elements. Therefore, it can be concluded that PP-DocBee-2B achieved the best performance in terms

Method	Printed text	Tables	Seals	Charts	Total Score
Closed-source MLMMs					
GPT-4o OpenAI (2024)	436	198	5	46	685
GLM-4V Flash GLM et al. (2024)	339	169	5	34	547
Open-source MLMMs					
InternVL2.5-2B Chen et al. (2024b)	363	182	4	47	596
Qwen2-VL-2B Wang et al. (2024)	476	167	8	29	680
PP-DocBee-2B	**517**	**202**	5	41	**765**

Table 3: Evaluation of existing OCR-Free MLLMs on internal Chinese benchmarks.

Method	DocVQA-val	ChartQA	InfoVQA-val	TextVQA	OCRBench	Internal-CN
Baseline	89.2	73.5	64.1	79.7	80.9	680
Baseline + 3.3M	89.6	74.3	65.0	80.6	81.6	726
Baseline + PPIDD	89.3	73.7	64.2	79.9	81.0	725
Baseline + 3.3M + PPIDD(12%)	89.6	74.6	65.0	80.6	81.6	743
Baseline + 3.3M + PPIDD(12%) + MS	89.7	74.6	65.0	80.6	81.4	745
Baseline + 3.3M + PPIDD(20%) + MS	89.7	74.6	65.2	80.6	80.7	765
Baseline + 3.3M + PPIDD(30%) + MS	89.7	74.4	64.8	80.6	80.9	761
PP-DocBee-2B	**90.1**	74.6	**65.4**	**81.2**	**82.8**	**765**

Table 4: Ablation study of PP-DocBee."3.3M" means adding 3.3M open source public datasets."PPIDD" means adding 477k synthetic dataset PPInfinityDocData. The number in the brackets after it, such as (20%), indicates the proportion of synthetic data to the total training data. If 3.3M public datasets and 477k PPInfinityDocData are added, the default proportion of synthetic data is 12%. This ratio adjustment represents the use of the Dynamic Ratio Sampling Training strategy."MS" means more steps, the default is 16k steps. Based on"Baseline + 3.3M + PPIDD + MS", PP-DocBee in the last row of the table also uses data preprocessing, which is to enlarge the image input scale.

of comprehensive accuracy. A more comprehensive and high-quality presentation of some cases can be found in Appendix A.

4.4 Ablation Study

To establish baseline performance, we initialized the Qwen2-VL-2B-Instruct model with pre-trained weights and extended its training using 330k samples from dvqa, chartqa, ai2d, docvqa, geoqa+ and synthdog. For comprehensive evaluation, we use five established English benchmark datasets as well as our proprietary internal Chinese scene annotation evaluation dataset (named Internal-CN). Table 4 presents a systematic ablation study evaluating the efficacy of two key innovations: Data Synthesis Strategy and Dynamic Ratio Sampling.

Effectiveness of Data Synthesis Strategy. To assess the impact of data synthesis strategies, we conducted a series of ablation experiments. These experiments compare the performance of the baseline model with the performance of models trained in different settings. We can see that on Internal-CN, the setting of"Baseline + PPIDD" (adding 477k synthetic data PPInfinityDocData to the baseline) can reach 725 points, while the setting of"Baseline + 3.3M" (adding 3.3M open source data to the baseline) can also reach 726 points. Because our Internal-CN evaluation set is closer to the real application scenario, it is closer to the source of the synthetic dataset, and has some dominance with the open source dataset. However, adding only 477k synthetic data has limited gains for five established English benchmark datasets. These results show that adding synthetic data (especially when tailored to specific weaknesses in the real dataset) can significantly improve the performance of the model's Chinese document scene understanding.

Effectiveness of Dynamic Ratio Sampling. We also conducted experiments to evaluate the effect of dynamic ratio sampling. After adding the 3.3M open source dataset and the 477k synthetic dataset, our training dataset contains English and Chinese document images from different fields, including financial reports, scientific articles, and legal documents. As can be seen in the overall table, the same setting of "Baseline + 3.3M + PPIDD + MS" is used, but the ratio of 20% is the best, with a maximum score of 765. Even if the ratio is increased to 30%, the score is only 761. This experimental result shows that assigning appropriate sampling weights to the training dataset according to the representativeness of the data and the characteristics of the evaluation set can significantly improve the ability of the model in some aspects.

5 Conclusion

PP-DocBee has made progress in the field of document image understanding through two major innovations: (1) a data synthesis strategy for Chinese document scenarios; (2) a set of training tricks, such as dynamic proportional sampling training, data preprocessing, OCR post-processing, etc. These innovations have greatly improved the model's ability to process and understand complex document images, setting a new benchmark in this field.

In future, we plan to conduct extensive experiments on a broader range of models and delve into more specialized areas, such as handwritten document analysis. We anticipate that the comprehensive solution offered by PP-DocBee will be well-equipped to tackle intricate document analysis tasks, thereby propelling the state-of-the-art in document image understanding.

References

aidapearson. Aida calculus math handwriting recognition dataset. https://www.kaggle.com/datasets/aidapearson/ocr-data, 2023.

PaddlePaddle Authors. Paddlemix, paddle multimodal integration and exploration. https://github.com/PaddlePaddle/PaddleMIX, 2023.

Azu. Handwritten-mathematical-expression-convert-latex. https://huggingface.co/datasets/Azu/Handwritten-Mathematical-Expression-Convert-LaTeX, 2023.

BAAI. Infinity-instruct dataset. https://huggingface.co/datasets/BAAI/Infinity-Instruct, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023b.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pp. 4291–4301, 2019.

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1511–1520, 2022.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv:2211.08545*, 2022.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024a.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.

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

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

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15138–15147, June 2023.

Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. pp. 1571–1576, 2019.

Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. pp. 845–855, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6904–6913, 2017.

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, Zhenchong Hu, Bo-Wen Zhang, Jijie Li, Dong Liang, Yingli Zhao, Yulong Ao, Yaoqi Liu, Fangxiang Feng, and Guang Liu. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data, 2024. URL https://arxiv.org/abs/2410.18558.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pp. 3608–3617, 2018.

Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. Icpr2018 contest on robust reading for multi-type web images. In *ICPR*, 2018.

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

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

Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Multi-scale adaptive cropping for multimodal large language models. *arXiv preprint arXiv:2408.02034*, 2024.

ift. Handwriting forms dataset. https://huggingface.co/datasets/ift/handwriting_forms, 2024.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDAR Workshops*, 2019.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pp. 5648–5656, 2018.

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

Kamizuru00. Diagram image to text dataset. https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text, 2024.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv:2203.06486*, 2022.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pp. 235–251, 2016.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pp. 4999–5007, 2017.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-Free document understanding transformer. In *ECCV*, 2022.

Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. Textstylebrush: transfer of text aesthetics from a single example. *IEEE Trans. PAMI*, 45(7):9122–9134, 2023.

LAION. Gpt-4v dataset. https://huggingface.co/datasets/laion/gpt4v-dataset, 2023.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv:2408.12637*, 2024.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv:2206.03001*, 2022.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv:2403.00231*, 2024a.

Zekun Li, Yijun Lin, Yao-Yi Chiang, Jerod Weinman, Solenn Tual, Joseph Chazalon, Julien Perret, Bertrand Duménieu, and Nathalie Abadie. ICDAR 2024 competition on historical map text detection, recognition, and linking. In *ICDAR*, 2024b.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *CVPR*, pp. 14963–14973, 2023.

Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv:2311.10774*, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 36, 2023d.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023e.

Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv:2407.15838*, 2024a.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023f.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023g.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv:2406.11833*, 2024b.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021a.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. 35:2507–2521, 2022a.

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

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv:2406.11069*, 2024.

Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, and Jean-Marc Ogier. Chic: Corporate document for visual question answering. In *ICDAR*, 2024.

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

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. pp. 2263–2279, 2022a.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL (Findings)*, pp. 2263–2279. Association for Computational Linguistics, 2022b.

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

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. In *WACV*, pp. 2199–2208. IEEE, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *WACV*, pp. 2582–2591. IEEE, 2022a.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. pp. 1697–1706, 2022b.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. pp. 1527–1536, 2020.

Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. pp. 947–952, 2019.

mychen76. Invoices and receipts ocr v1 dataset. https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1, 2024.

OleehyO. Latex formulas dataset. https://huggingface.co/datasets/OleehyO/latex-formulas, 2024.

OpenAI. Hello GPT-4o. https://openai.com/index/gpt-4v-system-card, 2024. Accessed: 2024-12-29.

parasam. Captcha dataset. https://www.kaggle.com/datasets/parsasam/captcha-dataset, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pp. 146–162, 2022.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *EMNLP*, 2015.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pp. 8317–8326, 2019a.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pp. 8317–8326. Computer Vision Foundation / IEEE, 2019b.

Hamed Rahimi Sujet AI, Allaa Boutaleb. Sujet-finance-qa-vision-100k: A large-scale dataset for financial document vqa. https://huggingface.co/datasets/sujet-ai/Sujet-Finance-QA-Vision-100k, 2024.

Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. pp. 1557–1562, 2019.

TAL. Tal open dataset. https://ai.100tal.com/dataset, 2023.

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. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. https://internvl.github.io/blog/2024-07-02-InternVL-2.0/, 2024.

TIGER-Lab. Visualwebinstruct dataset. https://huggingface.co/datasets/TIGER-Lab/VisualWebInstruct, 2024.

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

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. COCO-Text: Dataset and benchmark for text detection and recognition in natural images. *arXiv:1601.07140*, 2016.

Vicuna. Vicuna: An open chatbot impressing gpt-4. https://github.com/lm-sys/FastChat, 2023.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.

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

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, 2020.

Siwei Wu, Kang Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, Jiaheng Liu, Ruibo Liu, Xingwei Qu, Xuxin Cheng, et al. Mmra: A benchmark for multi-granularity multi-image relational association. *arXiv:2407.17379*, 2024.

Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *ECCV*, 2022.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv:2410.02712*, 2024.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv:2304.12244*, 2023.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap your own mathematical questions for large language models. *arXiv:2309.12284*, 2023a.

Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, et al. Icdar 2023 competition on structured text extraction from visually-rich document images. In *ICDAR*, 2023b.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv:2404.02078*, 2024.

Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34:509–521, 2019.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024.

Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. pp. 1577–1581, 2019.

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

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. Robut: A systematic study of table qa robustness against human-annotated adversarial perturbations. *arXiv:2306.14321*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *ACMMM*, 2022.

A APPENDIX

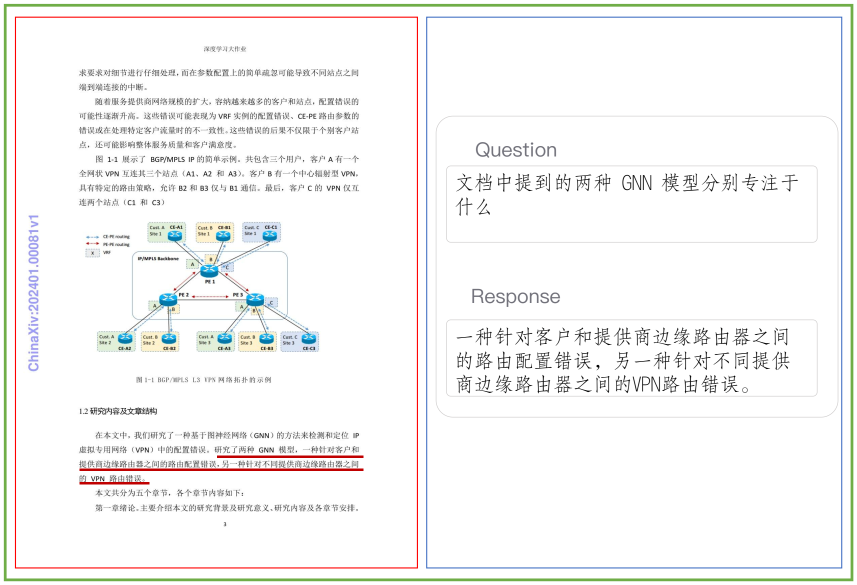


Figure 6: Case in science and engineering papers for document content understanding.

Figure 7: Case in science and engineering papers for table understanding.

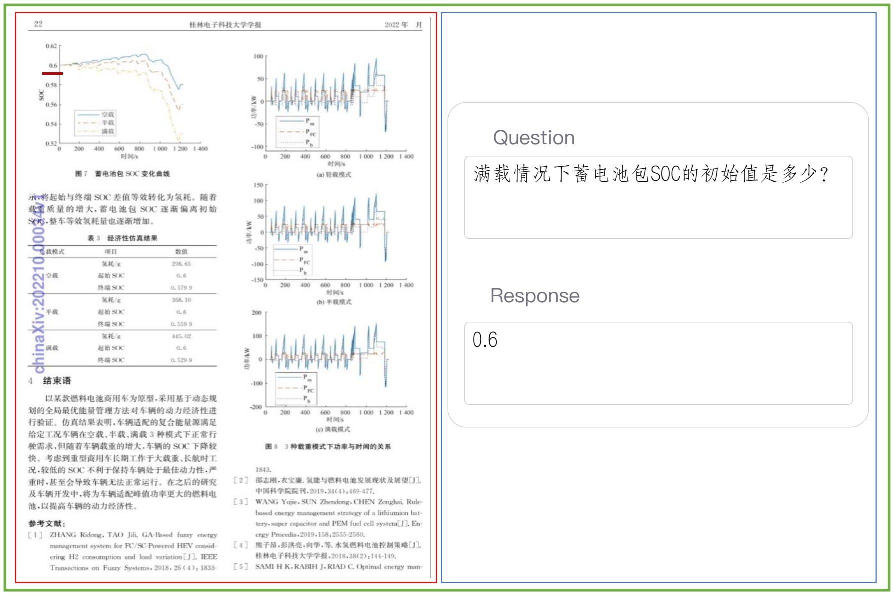


Figure 8: Case in science and engineering papers for chart understanding.

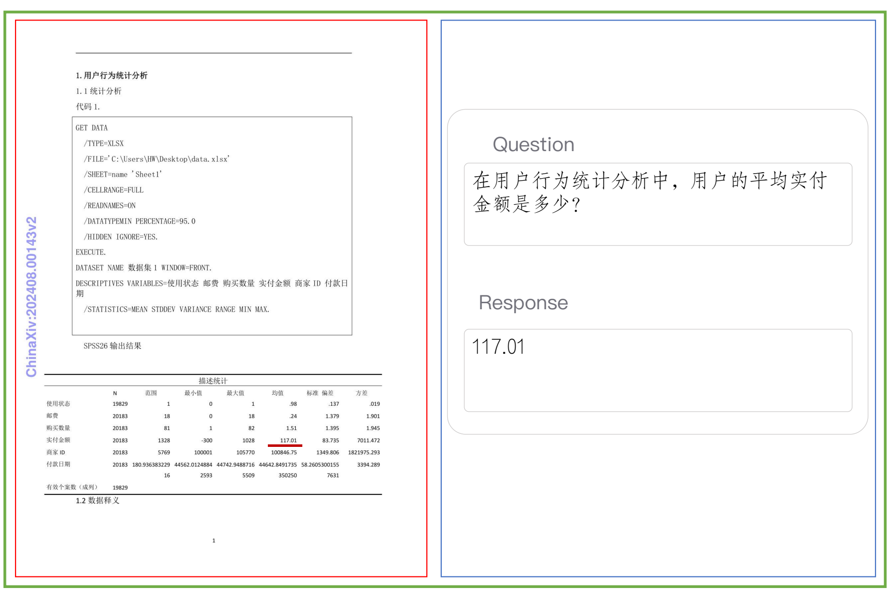


Figure 9: Case in financial reports for table understanding..