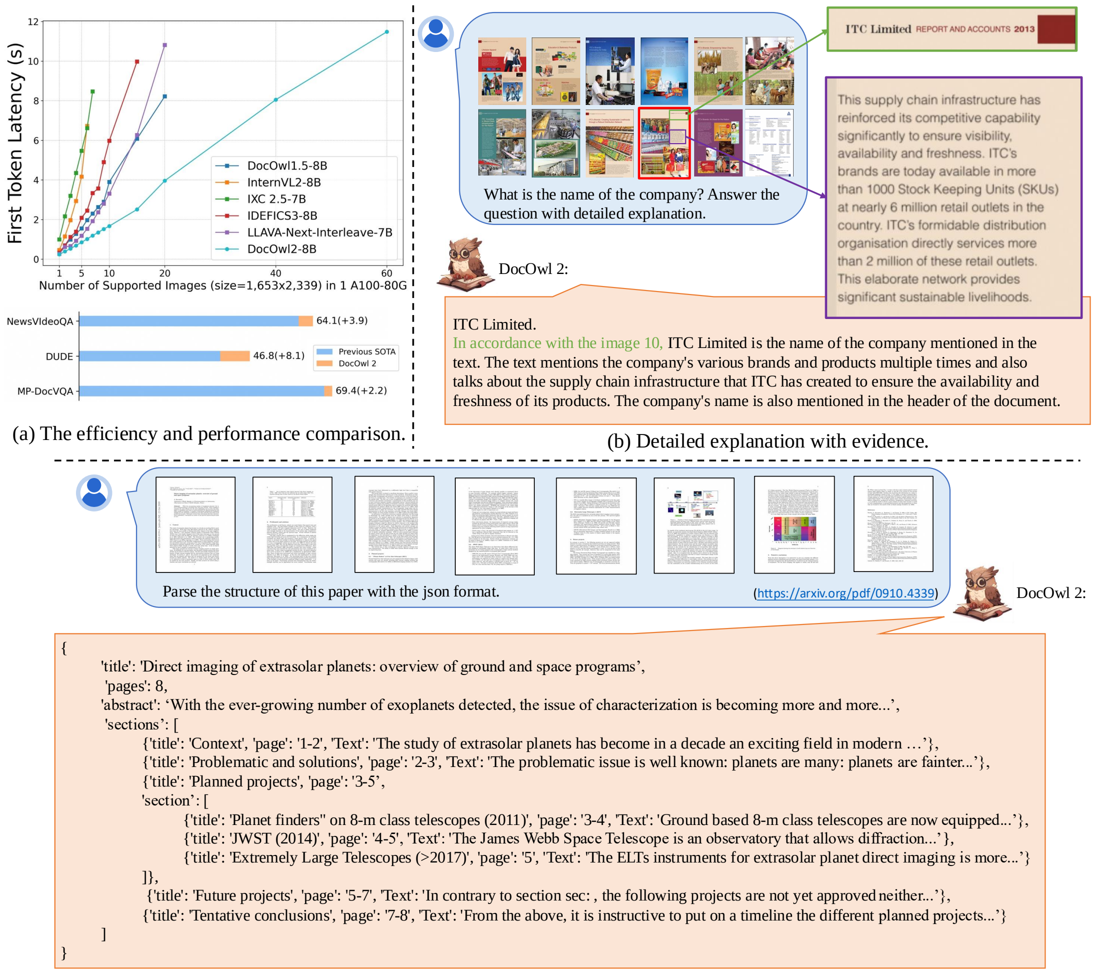
# mPLUG-DocOwl2: High-resolution Compressing for OCR-free Multi-page Document Understanding

Anonymous ACL submission




(a) The efficiency and performance comparison.

(b) Detailed explanation with evidence.

(c) Overall structure parsing for a document.

Figure 1: (a) mPLUG-DocOwl2 achieves state-of-the-art Multi-page Document Understanding performance with faster inference speed and less GPU memory; (b-c) mPLUG-DocOwl2 is able to provide a detailed explanation containing the evidence page as well as the overall structure parsing of the document.

## Abstract

Multimodel Large Language Models(MLLMs) have achieved promising OCR-free Document Understanding performance by increasing the supported resolution of document images. However, this comes at the cost of generating thousands of visual tokens for a single document image, leading to excessive GPU memory and slower inference times, particularly in multi-page document comprehension. In this work, to address these challenges, we propose a High-resolution DocCompressor module to compress each high-resolution document image into 324 tokens, guided by low-resolution global visual features. With this compression module, to strengthen multi-page document comprehension ability and balance both token efficiency and question-answering performance, we develop the DocOwl2 under a three-stage training framework: Single-image Pretraining, Multi-image Continue-pretraining, and Multi-task Finetuning. DocOwl2 sets a new state-of-the-art across multi-page document understanding benchmarks and reduces first token latency by more than 50%. Compared to single-image MLLMs trained on similar data, our DocOwl2 achieves comparable single-page understanding performance with less than 20% of the visual tokens. Our codes, models, and data will be publicly available.

## 1 Introduction

Understanding a multi-page document or news video is common in human daily life. To tackle such scenarios, Multimodal Large Language Models (MLLMs) (Ye et al., 2023c,d, 2024; Bai et al., 2023; Liu et al., 2023) should be equipped with the ability to understand multiple images with rich visually-situated text information. Different from natural images mainly comprising of objects, comprehending document images asks for a more fine-grained perception to recognize all texts. By encoding high-resolution document images with thousands of tokens, state-of-the-art Multimodal LLMs (Ye et al., 2023b; Hu et al., 2024; Chen et al., 2024; Dong et al., 2024a,b) achieves promising OCR-free document understanding performance, e.g., InternVL 2 (Chen et al., 2024) costs a average of 3k visual tokens for a A4-sized document page. However, as shown in Fig. 1(a), such long visual tokens not only result in long inference time but also occupy too much GPU memory, making it difficult to understand a complete document or video.

In this work, we argue that visual tokens of document images can be further compressed while maintaining both layout and most textual information. Existing compressing architecture in MLLMs are hard to balance information retention and token efficiency during document image encoding. As shown in Fig. 2(a), independently compressing each crop of a document image (Li et al., 2024b; Hu et al., 2024) could reduce visual tokens of each sub-image but still results in a long sequence of visual tokens after concatenating all sub-images. Leveraging learnable queries (Bai et al., 2023; Li et al., 2023a; Ye et al., 2023c) or selected tokens (Liu et al., 2024) as compressing guidance could produce an identical length of tokens for any resolution but overlook the overall layout information, as shown in Fig. 2(b). Layout-aware guidance is important for compressing visual features of document images because texts within a layout region are semantic-coherent and easier to summarize. For example, in a two-column paper, texts belonging to the 'Related Work' section are difficult to summarize with texts on the same line but belonging to the 'Method' section.

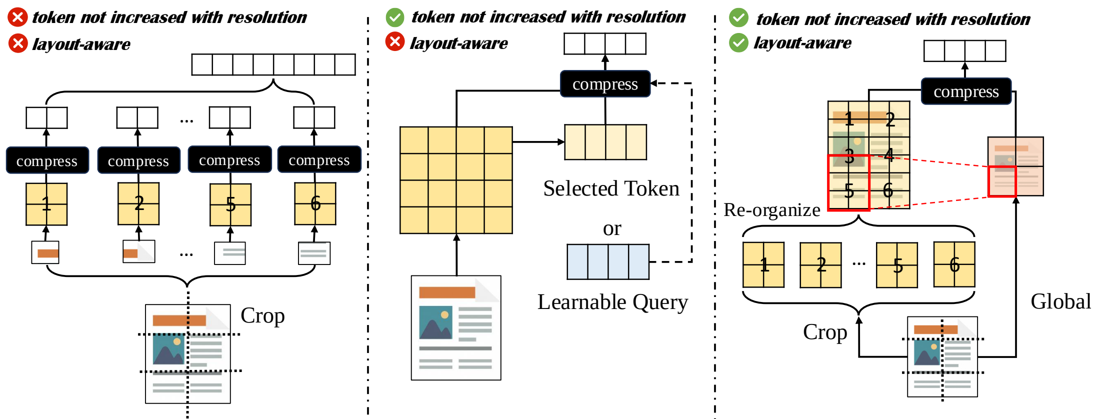
In this work, as shown in Fig. 2(c), we propose a layout-aware compressing architecture **High-resolution DocCompressor** based on cross-attention. Considering that a global low-resolution image can well capture the overall layout information, we utilize visual features of a global low-resolution image as the compressing guidance (query). Each visual feature in the global feature map just captures the layout information of partial regions. Therefore, each query attending to all high-resolution features will not only make information compression more difficult but also increase computation complexity. To summarize text information within a layout region, for each query from the global feature map, a group of high-resolution features with identical relative positions in the raw image is collected as compressing objects, sometimes spanning multiple sub-images. Besides, since the vision-to-text (V2T) module of MLLMs could convert visual features into textual feature space, we argue that compressing visual features after the vision-to-text module could better maintain textual semantics in document images. Therefore, based on the architecture of DocOwl 1.5 (Hu et al., 2024), we propose mPLUG-DocOwl2 by placing the High-resolution DocCompressor afther its V2T module: H-Reducer. To take full advantage of the compressing method, our model DocOwl2 is trained with a three-stage framework: Single-image Pretraining, Multi-image Continue-Pretraining, and Multi-task Finetuning to support both single-image and multi-image/frame understanding.

Our contributions in this work are three-fold:

- We propose a novel layout-aware compressing architecture to greatly reduce visual tokens of high-resolution document images.
- We design a three-stage training framework to empower DocOwl2 with both single-page and multi-page document understanding abilities.
- DocOwl2 achieves state-of-the-art performance on Multi-page Document understanding benchmarks with $< 50\%$ First Token Latency. Compared with state-of-the-art MLLMs with similar model size and training data, DocOwl2 achieves comparable performance with $< 20\%$ visual tokens on 10 single-image document benchmarks.

## 2 Related Work

**OCR-free Visual Document Understanding.** Visual Document Understanding aims to comprehend images with rich text information, including scans of document pages (Mathew et al., 2021; Tito et al., 2022; Landeghem et al., 2023; Zhang et al., 2023; Wei et al., 2023), infographics (Mathew et al., 2022), charts (Masry et al., 2022; Kafle et al., 2018;




(a) Compressing Each Crop (b) Guided by Learnable/Selected query (c) Ours: Layout-aware Compressing

Figure 2: Illustrations of different compressing methods for OCR-free document understanding.

Methani et al., 2020; Kahou et al., 2018), tables images (Pasupat and Liang, 2015; Chen et al., 2020; Zhong et al., 2020), webpage screenshots (Tanaka et al., 2021; Chen et al., 2021) and natural images with scene texts (Singh et al., 2019; Sidorov et al., 2020; Hu et al., 2021). Recently, many Multimodal Large Language Models have been proposed to perform visual document understanding in an OCR-free manner. mPLUG-DocOwl (Ye et al., 2023a) and UReader (Ye et al., 2023b) first propose to unify different tasks across 5 types of document images in the seq-to-seq format. To encode rich text information in high-resolution images, UReader (Ye et al., 2023b) proposes a Shape-adaptive Cropping Module to cut the raw image into multiple low-resolution sub-images and utilizes an identical low-resolution encoder to encode both sub-images and a global image. Monkey (Li et al., 2023b) proposes to employ a sliding window to partition high-resolution images and a resampler to reduce redundant information of each sub-image. mPLUG-DocOwl1.5 (Hu et al., 2024) increases the basic resolution of the low-resolution encoder and replaces the Visual Abstractor (Ye et al., 2023c) with 1 simple convolution layer to better maintain the structure information. DocPedia (Feng et al., 2023) directly processes high-resolution images in the frequency domain. CogAgent (Hong et al., 2023) proposes to utilize a high-resolution encoder to encode high-resolution visual features and a low-resolution encoder to encode low-resolution global features. Series work of InternLM-XComposer (Dong et al., 2024b,a) and InternVL (Chen et al., 2024) further optimize the cropping method or increase the cropping number and greatly improves the OCR-free Document Understanding performance. These works achieve promising performance but suffer from too many visual tokens for a high-resolution image (always $> 1k$ tokens for a common A4-sized document page), which hinders the development of OCR-free multi-page document understanding. Recent works explore enhancing the document understanding abilities of general-purpose MLLMs. However, they are either resource-intensive (Li et al., 2025) or can not achieve adorable performance (Kim and Seo, 2024). In contrast, we explore a more efficient model structure and training paradigm for multi-page document understanding. We believe these efforts can help to build models with more comprehensive capabilities.

**Visual Feature Compressing.** Reducing visual tokens of a single image enables a Multimodal Large Language Model with limited maximum sequence length to leverage more images as contexts to perform complex multimodal tasks, such as video understanding, embodied interaction, or multi-page document understanding. Some works (Zhang et al., 2024b; Shi et al., 2024; Li et al., 2024c) propose to ensemble and compress visual features from multiple vision encoders. For example, Eagle compresses visual features of 5 vision encoders to identical lengths of visual tokens and then fuses them by channel-level concatenation. Besides, there are also explorations to compress visual features of general images with fewer learnable queries, such as the Resampler (Alayrac et al., 2022; Bai et al., 2023), Abstractor (Ye et al., 2023c,d) and Q-former (Li et al., 2023a). Randomly initialized learnable queries can ensemble object information in general images but is hard to summarize rich text information in high-resolution document images. As a compromise solution, TokenPacker (Li et al., 2024b) proposes to compress each sub-image with its downsampled visual features as the query to perform cross-attention. To-

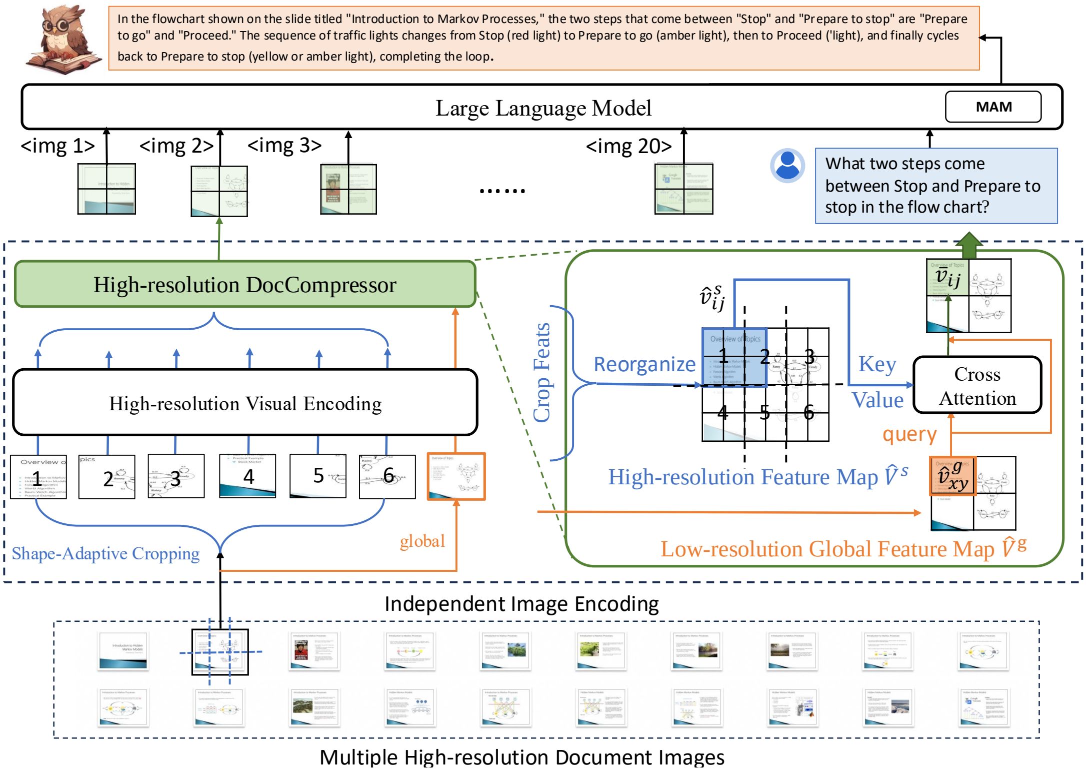



Figure 3: The architecture of DocOwl2. Each image is independently encoded by the pipeline of Shape-adaptive Cropping, High-resolution Visual Encoding and High-resolution DocCompressor.

kenPacker just reduces each sub-image's visual tokens, thus still creates more than 1k visual tokens when processing high-resolution document images. TextMonkey (Liu et al., 2024) first filters valuable visual tokens and then uses them as guidance to aggregate all visual tokens. Due to that valuable visual tokens are selected by measuring the token similarity, visual information of partial regions may not be covered and thus not well compressed during following cross-attention. In this work, our High-resolution DocCompressor leverages visual features from the row-resolution global images as the query, the ensembled feature map of sub-images as key and value. This not only produces a fixed number of visual tokens for images of any resolution but also covers all areas during compression. Compared to Mini-Gemini (Li et al., 2024c) which compresses general visual features, there are major two differences. Firstly, we make full use of global visual features and sub-image features produced by an identical low-resolution vision encoder and don't need to add an extra high-resolution encoder. Secondly, for better summarizing textual information in document images, our cross-attention is applied based on visual features that have been aligned with textual features of LLM. We argue that directly compressing outputs of the vision encoder loses semantic information while comprising features aligned with LLM is like summarizing texts and can better maintain textual semantics in document images. We conduct fair comparisons to support this hypothesis.

## 3 mPLUG-DocOwl2

As shown in Fig. 3, for multiple document images, DocOwl2 leverages a High-resolution Visual Encoding module and a High-resolution DocCompressor to encode each image independently. After that, a LLM is utilized for multimodal understanding.

### 3.1 High Resolution Vision Encoding

Following UReader (Ye et al., 2023b) and DocOwl 1.5 (Hu et al., 2024), DocOwl2 utilizes a parameter-free Shape-adaptive Cropping Module to preprocess high-resolution images. Concretely, it cuts each high-resolution image $I$ into $R \times C$ size-fixed sub-images $I^s = \{I^s_{xy}\}, 1 \leq x \leq R, 1 \leq y \leq C$, where cropping rows $R$ and columns $C$ are flexibly decided based on the raw resolution of $I$. Besides, to maintain the overall layout information, the raw image is also directly resized to a global image $I^g$.

After the cropping module, a low-resolution transformer-based vision encoder ViT (Dosovitskiy et al., 2021) is utilized to independently extract vision features of each sub-image and the global image as follows:

$$V^g = \mathrm{ViT}(I^g), \tag{1}$$

$$V^s_{xy} = \mathrm{ViT}(I^s_{xy}), 1 \le x \le R, 1 \le y \le C, \tag{2}$$

where both $V^g$ and $V^s_{xy}$ are visual features with the shape of $h \times w \times d$, $d$ is the feature dimension and $w, h$ are the width and height of the feature map.

Following DocOwl 1.5, after the ViT, for each sub-image or global image, we apply a vision-to-text module H-Reducer to ensemble horizontal 4 features by a convolution layer and align the feature dimension with the Large Language Model with a fully connected layer. The calculation of H-Reducer is represented as follows:

$$\hat{V} = \mathrm{FC}(\mathrm{Conv}(V)), \tag{3}$$

$$V \in \{V^g, V^s_{xy}\}, 1 \le x \le R, 1 \le y \le C, \tag{4}$$

where the shape of the visual feature map $\hat{V}$ is $h \times \frac{w}{4} \times \hat{d}$, $\hat{d}$ is the dimension of hidden states of the large language model.

### 3.2 High-resolution DocCompressor

A sentence/paragraph/document of text tokens can be compressed into fewer summary vectors while maintaining most semantics (Cheng et al., 2024; Ge et al., 2024; Chevalier et al., 2023). Besides, since visual features have been aligned with the textual feature space of large language models, the visual tokens of document images after the vision-to-text module can also be treated as textual tokens encoding different parts of textual information in the image. Thus, taking into account these two points, in this work, we argue that visually situated textual information of document images can also be further compressed into fewer tokens, especially after the vision-to-text alignment.

Texts from the same layout region are more appropriate to be fused into fewer tokens. After the vision-to-text module H-Reducer, the global visual feature $\hat{V}^g$ mainly encodes the overall text layout information while visual features of sub-images $\{\hat{V}^s_{xy}\}$ capture detailed textual information. Besides, due to both the global image and cropped sub-images come from an identical image, there is a clear mapping between the visual tokens of $\hat{V}^g$ and $\{\hat{V}^s_{xy}\}$. As shown in Fig. 3, each visual token in $\hat{V}^g$ can be aligned with $R \times C$ visual tokens in $\{\hat{V}^s_{xy}\}$. Therefore, we first re-organize feature maps of cropping images ($\{\hat{V}^s_{xy}\}, 1 \le x \le R, 1 \le y \le C$) to a complete feature map $\hat{V}^s$ according to their positions in the raw high-resolution image. Then, for each visual token in the feature map $\hat{V}^g$ of the global image, we collect its corresponding $R \times C$ visual tokens from $\hat{V}^s$ as the key and value, the cross-attention layer is calculated as follows:

$$\bar{v}_{ij} = \sigma\left(\frac{W^q \hat{v}^g_{ij} W^k \hat{v}^s_{ij}{}^T}{\sqrt{d_k}}\right) W^v \hat{v}^s_{ij} + \hat{v}^g_{ij}, \tag{5}$$

$$\hat{v}^g_{ij} \in \hat{V}^g, 1 \le i \le h, 1 \le j \le w/4, \tag{6}$$

$$\hat{v}^s_{ij} = [\hat{v}^s_{i'j'}] \subset \hat{V}^s, \tag{7}$$

$$(i-1)R + 1 \le i' \le iR, \tag{8}$$

$$(j-1)C + 1 \le j' \le jC, \tag{9}$$

where $\hat{v}^g_{ij}$ is a visual token from the global feature map and $\hat{v}^s_{ij}$ are visual tokens from the re-organized feature map of cropping images. $\hat{v}^g_{ij}$ and $\hat{v}^s_{ij}$ correspond to the same area in the raw image. $W^*$ are learnable matrics. $\sigma$ refers to softmax.

After high-resolution compressing, the compressed feature map of each image is organized into a sequence $\bar{V} = [\bar{v}_1, \bar{v}_2, ..., \bar{v}_{h \times \frac{w}{4}}]$ for subsequent understanding of the large language model.

### 3.3 Multi-image Modeling with LLM

Through the high-resolution compressing, the number of visual tokens for each high-resolution image is reduced from $(R \times C + 1) \times h \times \frac{w}{4}$ to $h \times \frac{w}{4}$. Such efficient vision encoding allows joint understanding of multiple document images with Large Language Models. To help the LLM better distinguish visual features from different images and understand the ordinal number of images, we add a textual ordinal token '<img $x$>' before the visual features of each image, where $x$ is the ordinal number. Overall, the decoding of the decoder for multiple images is as follows:

$$Y = \mathrm{LLM}([P_1; \bar{V}_1; P_2; \bar{V}_2, ..., P_n; \bar{V}_n; T]) \tag{10}$$

where $[;]$ means the concatenation operation, $n$ is the number of images, $P_x, 1 \le x \le n$ is the textual embedding of the ordinal token '<img $x$>', $\bar{V}_x$ is the visual features for each image, $T$ is the textual instruction and $Y$ is the predicted answer.

### 3.4 Model Training

DocOwl2 is trained with three stages: Single-image Pre-training, Multi-image Continue Pretraining, and Multi-task Finetuning.

At the first stage, to ensure the compressed visual tokens can encode most visual information,

Table 1: Comparison with OCR-free methods on single-image document understanding tasks. 'Token$^V$' means the average number of visual tokens of a single image. '**Bold**' means SOTA performance within the group and 'Underline' means achieving 80% SOTA performance among all baselines.

| Model | Domain | Size | Token$^V$ | Doc VQA | Info VQA | Deep Form | KLC | WTQ | Tab Fact | Chart QA | Text VQA | Text Caps | Visual MRC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IXC 2.5 | General | 7B | $\sim$ 5,118 | 90.9 | 69.9 | **71.2** | - | **53.6** | **85.2** | 82.2 | 78.2 | - | **307.5** |
| InternVL2 | General | 8B | $\sim$ 3,133 | **91.6** | **74.8** | - | - | - | - | **83.3** | **77.4** | - | - |
| DocOwl1.5 | Document | 8B | $\sim$ 1,698 | 82.2 | 50.7 | 68.8 | **38.7** | 40.6 | 80.2 | 70.2 | 68.6 | **131.6** | 246.4 |
| UReader | Document | 7B | $\sim$841 | 65.4 | 42.2 | 49.5 | 32.8 | 29.4 | 67.6 | 59.3 | 57.6 | 118.4 | **221.7** |
| TextMonkey | Document | 9B | 768 | 73.0 | 28.6 | 59.7 | **37.8** | 31.9 | - | 66.9 | 65.9 | - | - |
| TokenPacker | Document | 13B | $\sim$ 467 | 58.0 | - | - | - | - | - | - | - | - | - |
| QwenVL | General | 9B | 256 | 65.1 | 35.4 | - | - | - | - | 65.7 | 63.8 | - | - |
| DocOwl2 | Document | 8B | 324 | **80.7** | **46.4** | 66.8 | 37.5 | **36.5** | 78.2 | **70.0** | 66.7 | **131.8** | 217.4 |

especially visually situated texts, we first perform Unifed Structure Learning as DocOwl 1.5, which covers the learning of struct-aware document parsing, table parsing, chart parsing and natural image parsing of a single image.

After Single-image Pretraining, to empower our model with the ability to correlate multiple images, we further perform Multi-image Continue Pretraing with a struct-aware multi-page document parsing dataset MP-DocStruct1M. We design two symmetrical tasks of multi-image understanding: Multi-page Text Parsing and Multi-page Text Lookup. Given successive page images in a document, the Multi-page Text Parsing instructs the model to parse texts of specified one or two pages, such as `'Recognize texts in image 2 and image 10.'`. As for the Multi-page Text Lookup task, with texts from 1-2 pages as input, the model is required to predict the concrete ordinal number of images containing these texts, for example, `'Looking for the image with text <doc> ...</doc> and <doc> ...</doc>'`. Besides multi-image tasks, during this stage, we also randomly chose partial training samples from the first stage to avoid the catastrophic forgetting of structure parsing across different types of images.

Finally, we ensemble both single-page and multi-page instruction tuning datasets of document understanding to perform multi-task tuning. The task format includes concise question answering and detailed explanations.

The detailed introduction of training datasets of DocOwl2 can be found in Appendix A.1. More training details are introduced in Appendix A.2.

## 4 Experiments

### 4.1 Main Results

We compare DocOwl2 with state-of-the-art MLLMs on 10 single-image document understanding benchmarks, 2 Multi-page document Understanding benchmarks, and 1 text-rich video understanding benchmark. Both question-answering performance and the First Token Latency (seconds) are considered to show the effectiveness of our model.

Table 2: Comparison of performance and inference speed on DocVQA. 'FTL(s)' refers to the First Token Latency (seconds). 'IL(s)' refers to Instance Latency.

| Model | Size | Token$^V$ | FTL(s)↓ | IL(s)↓ | ANLS↑ |
|---|---|---|---|---|---|
| InternVL 2 | 8B | $\sim$ 3,198 | 0.94 | 2.46 | 91.6 |
| IXC 2.5 | 7B | $\sim$7,395 | 3.73 | 7.57 | 90.9 |
| DocOwl 1.5 | 8B | $\sim$1,806 | 0.58 | 1.84 | 82.2 |
| Idefics2 | 8B | 64 | **0.21** | **0.62** | 67.3 |
| Idefics2 | 8B | 320 | 0.89 | 2.15 | 74.0 |
| TextMonkey | 9B | 768 | 0.58 | 1.74 | 73.0 |
| DocOwl2 | 8B | 324 | 0.26 | 0.66 | 80.7 |

**Single-image Document Understanding** Compared with MLLMs (Ye et al., 2023b; Liu et al., 2024; Li et al., 2024b; Bai et al., 2023) with $< 1k$ visual tokens, our DocOwl2 achieves better or comparable performance on 10 benchmarks. Especially, with fewer visual tokens, our model outperforms both TextMonkey (Liu et al., 2024) and TokenPacker (Li et al., 2024b) which also aim to compress visual tokens, showing that our layout-aware architecture High-resolution DocCompressor is better at summarizing and maintaining textual information in high-resolution document images. Besides, compared with state-of-the-art MLLMs (Dong et al., 2024b; Chen et al., 2024; Hu et al., 2024) with $> 1k$ visual tokens, DocOwl2 achieves $> 80\%$ performance on 7/10 benchmarks while with $< 20\%$ visual tokens. A more comprehensive comparison with existing OCR-free models can be found in Appendix B.1.

Furthermore, we compare the First Token Latency (seconds) on the most frequently compared dataset DocVQA (Mathew et al., 2021). As shown in Table 2, the far greater number of visual tokens enable InternVL 2 (Chen et al., 2024) and IXC 2.5 (Dong et al., 2024b) to achieve better perfor-

Table 3: The OCR-free performance comparison on multi-page/video document understanding benchmarks. 'FTL(s)' refers to the First Token Latency. 'Token$^{V}$' means the average number of visual tokens of a single page/frame. LLaVA-Next-Interleave-7B*: fine-tuned with the same data of DocOwl2 for held-in evaluation.

| **Model** | **Token$^{V}$** | **MP-DocVQA** | | **DUDE** | | **NewsVideoQA** | |
|---|---|---|---|---|---|---|---|
| | | FTL(s)↓ | ANLS↑ | FTL(s)↓ | ANLS↑ | FTL(s)↓ | ANLS↑ |
| LongVA-7B | ∼2,029 | 2.13 | 60.80 | 2.26 | 38.37 | 4.29 | 50.61 |
| Idefics3-8B | ∼838 | 2.26 | 67.15 | 2.29 | 38.65 | 6.39 | 60.16 |
| LLaVA-Next-Interleave-7B | 729 | 1.56 | 44.87 | 1.47 | 28.03 | 4.35 | 56.66 |
| LLaVA-Next-Interleave-7B* | 729 | 1.56 | 49.99 | 1.47 | 39.02 | 4.35 | 62.38 |
| DocOwl2-8B | 324 | **0.95** | **69.42** | **0.94** | **46.77** | **1.17** | **64.09** |

Table 4: Ablation study about the architecture of the compressor on single-image document benchmarks. 'Img$^{base}$' refers to the basic resolution of the global image and each sub-image.

| | **Img$^{base}$** | **Crop** | **Compressor** | | | | | **DocVQA** | **WTQ** | **ChartQA** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Name | Compressing | Layer | Position | Token$^{V}$ | | | |
| r1 | 448 | 9 | Resampler | learnable query | - | after H-Reducer | 256 | 69.0 | 29.4 | 66.6 |
| r2 | 448 | 9 | CAbstractor | Adaptive Mean | - | after H-Reducer | 256 | 73.0 | 32.6 | 67.6 |
| r3 | 448 | 9 | DocCompressor | Group Att | 2 | after H-Reducer | 256 | 76.1 | 35.1 | 69.2 |
| r4 | 448 | 9 | DocCompressor | Group Att | 2 | after ViT | 256 | 75.7 | 33.3 | 68.7 |
| r5 | 448 | 9 | DocCompressor | Complete Att | 2 | after H-Reducer | 256 | 74.4 | 33.7 | 68.2 |
| r6 | 448 | 9 | DocCompressor | Group Mean | - | after H-Reducer | 256 | 74.6 | 31.9 | 68.2 |
| r7 | 448 | 9 | DocCompressor | Group Att | 1 | after H-Reducer | 256 | 76.4 | 34.2 | 69.2 |
| r8 | 448 | 9 | DocCompressor | Group Att | 4 | after H-Reducer | 256 | 75.9 | 35.8 | 70.1 |
| r9 | 448 | 12 | DocCompressor | Group Att | 2 | after H-Reducer | 256 | 76.8 | 35.6 | 69.5 |
| r10 | 504 | 12 | DocCompressor | Group Att | 2 | after H-Reducer | 324 | 78.7 | 36.7 | 69.4 |

mance but also result in higher inference time. Considering the model architecture and training data, it's most fair to compare DocOwl2 with DocOwl 1.5. After adding the High-resolution DocCompressor, with similar training data of OCR learning, DocOwl2 achieves 98% performance of DocOwl 1.5 while reducing 50% First Token Latency with just 20% visual tokens, validating the effectiveness of our compressor for compressing visually-situated text information. Similar comparisons on more benchmarks can be found in Appendix B.1.

**Multi-page/Video Document Understanding** In such benchmarks, we choose recently proposed Multimodal LLMs (Zhang et al., 2024a; Laurençon et al., 2024; Li et al., 2024a) with multi-page OCR-free document understanding abilities and can be fed into more than 10 images under a single A100-80G as baselines. As shown in Table 3, with fewer visual tokens for a single image/frame, DocOwl2 achieves better question-answering performance and much less First Token Latency, validating the good balance of DocOwl2 between the OCR-free document understanding performance and token efficiency.

## 4.2 Ablation Study

**Compressor Architecture.** We compare different compressing architectures with an identical training pipeline of Single-image Pretraing and Single-image Document Understanding Finetuning, keeping both training data and setting consistent.

As shown in Table 4, compared with CAbstractor (Cha et al., 2023), the Resampler (Bai et al., 2023) achieves worse document understanding performance (r2 vs r1). This shows that due to no prior knowledge, such as spatial relationship, is leveraged as compressing guidance, utilizing queries learned from scratch to compress rich visually-situated text information is more challenging than simple adaptive mean pooling. Our High-resolution DocCompressor outperforms CAbstractor (r3 vs r2), validating that leveraging global visual features as layout-aware guidance can better distinguish the information density of each fine-grained visual feature and therefore maintain more visually-situated text information.

Instead of placing the compressor after the vision-to-text module H-Reducer, we also try inserting it between the vision encoder and the vision-to-text module. Such a setting results in performance decreases across three datasets (r4 vs r3), validating our hypothesis that compressing features after the vision-to-text module is like summarizing textual features and can maintain more textual semantics while compressing visual features after the visual encoder loses more visually situated text information. Besides, without aligning each query token in the global feature map with $R \times C$ fine-grained visual tokens from the re-organized feature map to perform attention within a group as Eq. (9), we

Table 5: Ablation study about the training stages of DocOwl2. 'Page Num' and 'Evidence Page' refer to the number of input page images and the page ordinal number with the ground-truth answer.

| | Pretraining | | SFT | | DocVQA | MP-DocVQA | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Single Image | Multi Image | Single Image | Multi Image | | Page Num 1 | Page Num 2-10 | Page Num >10 | Evidence Page 1 | Evidence Page 2-10 | Evidence Page >10 | Overall |
| r1 | ✓ | | ✓ | | 78.7 | 81.3 | 55.0 | 5.8 | 67.7 | 45.9 | 6.2 | 54.2 |
| r2 | ✓ | | | ✓ | 75.2 | 78.7 | 65.2 | 34.6 | 74.3 | 54.9 | 40.9 | 63.8 |
| r3 | ✓ | ✓ | | ✓ | 74.2 | 78.9 | 65.7 | 37.9 | 74.2 | 56.8 | 43.4 | 64.7 |
| r4 | ✓ | ✓ | ✓ | ✓ | **80.7** | **83.3** | **70.2** | **42.5** | **78.6** | **60.9** | **53.6** | **69.4** |

try utilizing each query token to attend all visual tokens of sub-images. Such complete attention not only brings higher computational complexity but also causes performance decreases (r5 vs r3), showing that the positional correspondence between the global visual map and the re-organized fine-grained visual map is a reliable prior knowledge for compressing visual features efficiently. Furthermore, directly performing mean pooling on each group of $R \times C$ fine-grained visual features underperforms utilizing global visual features as the query (r6 vs r3), proving the importance of reliable guidance during compressing.

Compared with 2 layers of cross-attention, decreasing cross-attention layers bring a slight performance increase on DocVQA (Mathew et al., 2021) but more performance decrease on WikiTablesQA (WTQ) (Pasupat and Liang, 2015) (r7 vs r3). Further increasing to 4 layers doesn't significantly improve performance (r8 vs r3). This shows that compressing high-resolution visual features doesn't require a deep neural network. Finally, increasing the maximum number of crops and the base resolution of the global image or each sub-image are two main strategies to increase the supported input resolution. Our experiments show that increasing the cropping number (r9 vs r3) or basic resolution (r10 vs r9) benefits the document understanding performance. Increasing basic resolution brings more improvement because of more visual tokens after compressing.

**Three-stage Training.** DocOwl2 is trained with three stages: Single-image Pretraining, Multi-image Continue-pretraining, and Multi-task Finetuning. Table 5 shows the influence of each stage for OCR-free single-page and multi-page document understanding. With the Single-image Pretraining and Single-image finetuning (r1), the model achieves promising performance on single-page benchmark DocVQA and documents from MP-DocVQA with only 1 page. Although only trained with 1 image as the input, the model can also achieve around 50% accuracy when fed into 2-10 page images. However, the model struggles to understand documents with more than 10 pages, which greatly exceeds the number of input images during training and brings great difficulty in correlating images and finding answers. Performing Multi-image Fintuning could greatly improve the model's ability to understand multiple images (r2 vs r1). Furthermore, adding the Multi-image Continue-pretraining could also improve the question-answering performance on downstream datasets, especially for documents with more than 10 pages (r3 vs r2). This demonstrates that parsing texts of the specified page or judging which pages contain specified texts among multi-page documents is a basic ability for multi-page document understanding. Finally, by ensembling both single-image and multi-image instruction tuning sets (r4), DocOwl2 achieves the best performance on both single-page and multi-page document benchmarks, showing the cross-improvement between single-image and multi-image comprehension.

Qualitative results of multi-page text parsing, text lookup, question answering with detailed explanations can be found in Appendix B.5.

## 5 Conclusion

We propose DocOwl2, a Multimodal LLM for efficient OCR-free Multi-page Document Understanding. The novel architecture High-resolution DocCompressor compresses each high-resolution document image into 324 tokens through cross-attention with the global visual feature as guidance, and reorganized features of cropped images as keys and values. A carefully designed three-stage training framework empowers the model with multi-page understanding ability and maintains single-page performance after compressing visual tokens. With fewer visual tokens, DocOwl2 outperforms existing compressing methods on single-page document understanding benchmarks, and achieves OCR-free state-of-the-art performance on two multi-page document understanding benchmarks and 1 text-rich video understanding benchmark.

## 6 Limitation

In this work, we propose a compressing architecture High-resolution DocCompressor for reducing visual tokens of high-resolution document images. Due to the compressor being placed between the vision-to-text module and the LLM, extra training for compressing visual tokens and re-aligning with LLM is indispensable. A more efficient method of compressing visual tokens and reduce training costs for re-aligning with LLMs can better leverage existing MLLMs, which is left as future work.

## 7 Ethics Statement

Initialized from a general Multimodal Large Language Model trained with massive web data, DocOwl2 may also suffer from issues of LLMs such as toxic language and bias (Bender et al., 2021). However, the three-stage training in this work focuses on parsing texts or questioning answering for publicly available document images. This introduces few biases relevant to ethical issues.

## References


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

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

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal LLM. *CoRR*, abs/2312.06742.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact : A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.

Xingyu Chen, Zihan Zhao, Lu Chen, JiaBao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. Websrc: A dataset for web-based structural reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4173–4185.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *CoRR*, abs/2404.16821.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. *CoRR*, abs/2405.13792.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *EMNLP*, pages 3829–3846. Association for Computational Linguistics.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024a. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k HD. *CoRR*, abs/2404.06512.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024b. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k HD. *CoRR*, abs/2404.06512.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net.

Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2023. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *CoRR*, abs/2311.11810.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *ICLR*. OpenReview.net.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogagent: A visual language model for GUI agents. *CoRR*, abs/2312.08914.

Anwen Hu, Shizhe Chen, and Qin Jin. 2021. Question-controlled text-aware image captioning. In *ACM Multimedia*, pages 3097–3105. ACM.

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

Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2023. Watching the news: Towards videoqa models that can read. In *WACV*, pages 4430–4439. IEEE.

Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: understanding data visualizations via question answering. In *CVPR*, pages 5648–5656. Computer Vision Foundation / IEEE Computer Society.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. In *ICLR (Workshop)*. OpenReview.net.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *ECCV (28)*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.

Geewook Kim and Minjoon Seo. 2024. On efficient language and vision assistants for visually-situated natural language understanding: What matters in reading and reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16978–17000, Miami, Florida, USA. Association for Computational Linguistics.

Jordy Van Landeghem, Rafal Powalski, Rubèn Tito, Dawid Jurkiewicz, Matthew B. Blaschko, Lukasz Borchmann, Mickaël Coustaty, Sien Moens, Michal Pietruszka, Bertrand Anckaert, Tomasz Stanislawek, Pawel Józiak, and Ernest Valveny. 2023. Document understanding dataset and evaluation (DUDE). In *ICCV*, pages 19471–19483. IEEE.

Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pages 87874–87907. Curran Associates, Inc.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 18893–18912. PMLR.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jianke Zhu, and Lei Zhang. 2024b. Token-packer: Efficient visual projector for multimodal LLM. *CoRR*, abs/2407.02392.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024c. Mini-gemini: Mining the potential of multi-modality vision language models. *CoRR*, abs/2403.18814.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023b. Monkey: Image resolution and text label are important things for large multi-modal models. *CoRR*, abs/2311.06607.

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

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024. Textmonkey: An ocr-free large multimodal model for understanding document. *CoRR*, abs/2403.04473.

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

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

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

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1516–1525. IEEE.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *ACL (1)*, pages 1470–1480. The Association for Computer Linguistics.

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

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *ECCV (2)*, volume 12347 of *Lecture Notes in Computer Science*, pages 742–758. Springer.

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

Tomasz Stanislawek, Filip Gralinski, Anna Wróblewska, Dawid Lipinski, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemyslaw Biecek. 2021. Kleister: Key information extraction datasets involving long documents with complex layouts. In *ICDAR (1)*, volume 12821 of *Lecture Notes in Computer Science*, pages 564–579. Springer.

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

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *AAAI*, pages 13878–13888. AAAI Press.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2022. Hierarchical multimodal transformers for multi-page docvqa. *CoRR*, abs/2212.05935.

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

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023. Vary: Scaling up the vision vocabulary for large vision-language models. *CoRR*, abs/2312.06109.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, Shiyang Feng, Bin Wang, Chao Xu, Conghui He, Pinlong Cai, Min Dou, Botian Shi, Sheng Zhou, Yongwei Wang, Bin Wang, Junchi Yan, Fei Wu, and Yu Qiao. 2024. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *CoRR*, abs/2406.11633.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *CoRR*, abs/2307.02499.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. 2023b. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *EMNLP (Findings)*, pages 2841–2858. Association for Computational Linguistics.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *Preprint*, arXiv:2408.04840.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023c. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023d. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257.

Liang Zhang, Anwen Hu, Jing Zhang, Shuo Hu, and Qin Jin. 2023. Mpmqa: multimodal question answering on product manuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13958–13966.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *CoRR*, abs/2406.16852.

Ruiyi Zhang, Yufan Zhou, Jian Chen, Jiuxiang Gu, Changyou Chen, and Tong Sun. 2024b. Llava-read:

Enhancing reading ability of multimodal language models. *Preprint*, arXiv:2407.19185.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno-Yepes. 2020. Image-based table recognition: Data, model, and evaluation. In *ECCV (21)*, volume 12366 of *Lecture Notes in Computer Science*, pages 564–580. Springer.

# A Model Training

## A.1 Training Data

We utilize DocStruct4M (Hu et al., 2024) as the training data of the first stage.

In the second stage, we construct training samples of Multi-page Text Parsing and Multi-page Text Lookup based on partial documents from two datasets of PixParse[1][2].

As for the third Multi-task Tuning stage, we leverage DocDownstream-1.0 (Hu et al., 2024) and DocReason25K (Hu et al., 2024) as single-image datasets. DocDownstream-1.0 is an ensembled dataset comprising of DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), DeepForm (Svetlichnaya, 2020), KLC (Stanislawek et al., 2021), WTQ (Pasupat and Liang, 2015), TabFact (Chen et al., 2020), ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), TextCaps (Sidorov et al., 2020) and VisualMRC (Tanaka et al., 2021). DocReason25K is a question-answering dataset with detailed explanations. As for multi-image understanding, we ensemble 2 document datasets, MP-DocVQA (Tito et al., 2022) and DUDE (Landeghem et al., 2023), and 1 news video dataset NewsVideoQA (Jahagirdar et al., 2023) as concise question-answering datasets. MP-DocVQA contains 46k question-answering pairs on 60k page images scanned from 6k industry documents with rich tables, diagrams, pictures, and both handwritten and printed texts. DUDE covers more domains of documents, including medical, legal, technical, financial, etc. It contains 41k question-answering pairs on 5k documents. NewsVideoQA collects news videos with rich visually-situated texts from diverse English news channels around the world, such as BBC, CNN, etc. It contains 8k question-answering pairs framed on 3k videos. Besides, to trigger the ability of detailed explanations with evidence pages, we built MP-DocReason51K based on DocReason25K. Concretely, for each single-image sample from DocReason25K, we construct two multi-image samples with noisy images randomly chosen from the same or different categories. After randomly inserting the evidence image into noisy images, we add an extra evidence description (e.g., `'According to the 5th image,'`) into the raw detailed explanation to get the target of multi-image samples. Most question-answering samples just focus on 1-2 pages of a document, to further strengthen the ability of a comprehensive understanding of a document, we leverage a small part of annotations from DocGenome (Xia et al., 2024) to construct text sequences in the JSON format, which represents the hierarchical structure of a scientific paper and partial detailed texts.

Table 6 shows the detailed statistic of training data at each stage.

## A.2 Implementation Details

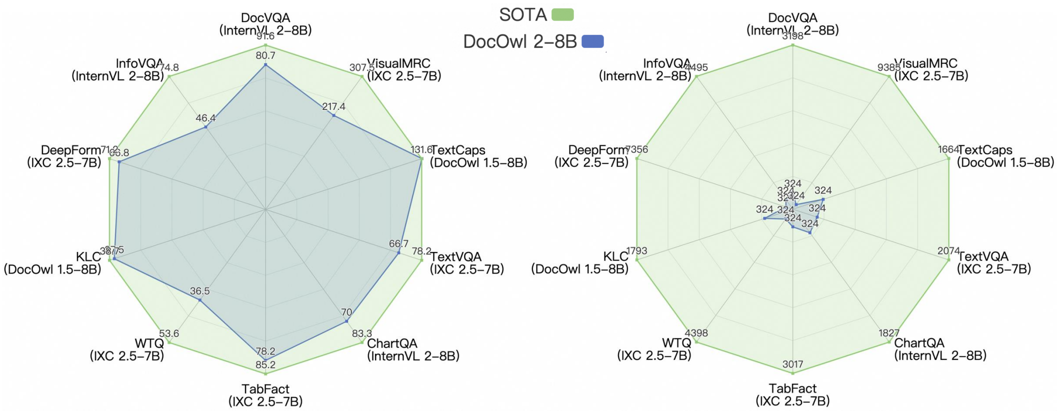
The maximum number of crops is set to 12. The resolution of each sub-image or the global image is 504x504. The High-resolution DocCompressor comprises of 2 layers of cross attention. Initialized from mPLUG-Owl2 (Ye et al., 2023d), the vision encoder (ViT/L-14 (Dosovitskiy et al., 2021)), H-Reducer and High-resolution DocCompressor are trained during the Sinlge-image Pretraining. Besides, the main parameters of the Large Language Model (Touvron et al., 2023) are frozen while a Modality Adaptive Module (MAM) (Ye et al., 2023d) used to distinguish visual and textual features in the LLM is tuned. The first stage takes 12k steps on 32 A100 GPUs for 84 hours with a batch size of 1,024 and a learning rate of 1e-4. During the Multi-image Continue-pretraining, the vision encoder is further frozen and the H-Reducer, High-resolution DocCompressor and MAM is tuned. The second stage takes 2.4k steps on 32 A100 GPUs in 130 hours with a batch size of 1,024 and the learning rate set as 2e-5. At the final Multi-task Finetuning stage, all parameters except the vision encoder are optimized. The batch size, training step, and learning rate at this stage are set as 256, 9k, 2e-5 respectively. This training stage takes 125 hours to converge with 32 A100 GPUs.

[1]https://huggingface.co/datasets/pixparse/idl-wds

[2]https://huggingface.co/datasets/pixparse/pdfa-eng-wds

Table 6: Detailed statistic of training datasets of DocOwl2.

| Training Stage | Input Image | Dataset | Num |
|---|---|---|---|
| Single-image Pretraining | Single | DocStruct4M | 4,036,402 |
| Multi-image Continue Pretraining | Single | DocStruct4M | 501,781 |
| | Multiple | MP-DocStruct1M | 1,113,259 |
| Multi-task Finetuning | Single | DocVQA, InfoVQA, DeepForm, KLC, WTQ, TabFact, ChartQA, TextVQA, TextCaps, VisualMRC | 552,315 |
| | | DocReason25K | 25,877 |
| | Multiple | MP-DocVQA | 70,154 |
| | | DUDE | 35,438 |
| | | NewsVideoQA | 8,619 |
| | | MP-DocReason51K | 51,754 |
| | | DocGenome12K | 12,010 |




(a) Performance

(b) Average Number of Visual Tokens

Figure 4: The comparison of our DocOwl2 with state-of-the-art Multimodal Large Language Models on (a) OCR-free performance and (b) the average number of visual tokens on 10 Visual Document Understanding benchmarks.

## B Experiments

### B.1 Single-image Document Understanding

We divide baselines into three groups: (a) models without Large Language Models as decoders (Kim et al., 2022; Lee et al., 2023), (b) Multimodal LLMs (Hong et al., 2023; Dong et al., 2024b; Chen et al., 2024; Li et al., 2024b; Hu et al., 2024; Feng et al., 2023; Li et al., 2023b) with an average number of visual tokens over 1k for a single document image and (c) Multimodal LLMs (Ye et al., 2023a,b; Liu et al., 2024; Li et al., 2024b; Bai et al., 2023) with an average number of visual tokens less than 1k. As shown in Table 8, although specifically fine-tuned on each downstream dataset, Donut (Kim et al., 2022) or PixsStruct (Lee et al., 2023) are not as good as Multimodal LLMs, showing the potential of MLLMs for generalized OCR-free document understanding. Among models with $<$1k visual tokens, DocOwl2 achieves state-of-the-art performance. Compared with MLLMs with $>$1k visual tokens, DocOwl2 achieves $> 80\%$ performance on 7 benchmarks while with $< 20\%$ visual tokens. Fig. 4 visualizes the comparison with SOTA in terms of question-answering performance and the number of visual tokens.

Table 9 further shows the performance and inference speed comparison on the 3 most frequently compared benchmarks, representing document, chart, and natural images.

### B.2 DocCompressor with different models

Our proposed DocCompressor is theoretically compatible with most MLLMs that have a vision-to-text module. To verify this, we insert DocCompressor into the LLaVA-Next-Interleave between its vision-to-text MLP and LLM. We finetune both the original model and model with DocCompressor with the same data of DocOwl2 and evaluate across both single-page and multi-page document understanding benchmarks. As shown in Table 10, LLaVA-Next-Interleave (w/ doccompressor) achieves comparable performance with LLaVA-Next-Interleave with less visual tokens. It validates that our compression module can be applied with a different backbone model.

### B.3 DocCompressor versus Mini-Gemini

Though Mini-Gemini (Li et al., 2024c) also explore to mixing low and high-resolution features via cross-attention, there are two major difference between DocCompressor and Mini-Gemini. First, DocCompressor uses a single vision encoder combined with image cropping to encode high resolution images, and Mini-Gemini relies on an additional high resolution encoder. Second, our DocCompressor merges high-resolution information after the vision-to-text module, and Mini-Gemini does it before. To show the advantages of our framework, we train both structure with the same training recipe to make a fair comparison. As shown in Table 7, our DocCompressor outperforms Mini-Gemini on all document understanding benchmarks, which verifies the effectiveness of the design of our DocCompressor.

Table 7: Comparison between Mini-Gemini and DocCompressor over document understanding benchmarks.

| Structure | DocVQA | WTQ | ChartQA | InfoVQA | DeepForm | KLC |
|---|---|---|---|---|---|---|
| Mini-Gemini | 75.7 | 33.3 | 68.7 | 41.6 | 58.4 | 37.0 |
| DocCompressor | **76.1** | **35.1** | **69.2** | **41.7** | **59.5** | **37.5** |

### B.4 Text Capacity Analysis of the Visual Embedding

To analyze the text capacity of the visual embedding, we further synthesize several A4-sized document images with different font sizes and numbers of characters to examine the parsing performance of DocOwl2 with 324 visual tokens. Concretely, we create an A4-sized document page with the resolution of $595 \times 842$ through the PyMuPDF and fill it with font sizes from 10 to 20 of English texts collected from a Wikipedia to synthesize multiple document pages. The number of characters ranges from 1,540 to 6,104. We let DocOwl2 parse the word inside these images and evaluate the result by ANLS score. Fig. 5 shows that DocOwl2 could almost perfectly parse a document with a document less than 5,000 characters. We shows a decline ANLS score when the character numbers exceeds 5,000, but it still maintain an ANLS score of 92.56% given font size of 10, which contains 6,104 characters or 1,502 tokens inside a A4-sized page. This result demonstrated that our model has strong text capacity with only 324 visual tokens.

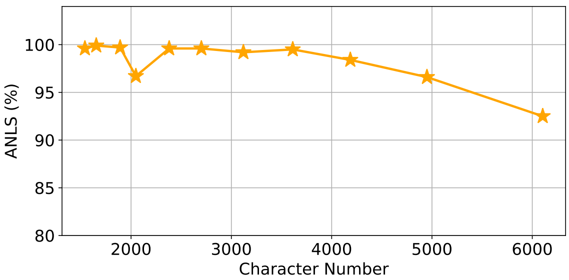



Figure 5: Parsing performance on A4-sized document image.

### B.5 Qualitative Results

As shown in Fig. 6, after the Multi-image Continue Pretraining stage, DocOwl2 is able to locate the corresponding image of the given texts accurately. Besides, although representing each high-resolution image with just 324 tokens, DocOwl2 is still capable of parsing detailed texts of specified two images, validating the promising OCR-free multi-page document understanding performance of DocOwl2 . It also demonstrates our proposal that 324 tokens are enough to encode detailed text information in common A4-sized document pages and the effectiveness of our High-resolution DocCompressor.

After the Multi-task Finetuning, given multiple images and a question, DocOwl2 can give a simple answer first and then provide a detailed explanation with the evidence, as shown in Fig. 7. DocOwl2 can comprehend not only page images rendered from PDF files (Fig. 7(c)) but also scan images of a document (Fig. 7(a-b)). When a question is unanswerable, DocOwl2 can also tell and give corresponding reasons (Fig. 7(c)).

Besides multi-page documents, DocOwl2 is also capable of understanding text-rich videos. As shown in Fig. 8, among similar frames within a video, DocOwl2 can distinguish fine-grained textual differences, locate relevant frames, and give accurate answers.

Table 8: Comparison with OCR-free methods on single-image document understanding tasks. The '$*$' refers to models without LLMs and separately fine-tuned on each downstream task. 'Token$^V$' means the average number of visual tokens of a single image. '**Bold**' means SOTA performance within the group and 'Underline' means achieving 80% SOTA performance among all baselines.

| | Model | Size | Token$^V$ | Doc VQA | Info VQA | Deep Form | KLC | WTQ | Tab Fact | Chart QA | Text VQA | Text Caps | Visual MRC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Donut$^*$ | <1B | 4,800 | 67.5 | 11.6 | 61.6 | 30.0 | 18.8 | 54.6 | 41.8 | 43.5 | 74.4 | 93.91 |
| | Pix2Struct$^*_{base}$ | <1B | 2,048 | 72.1 | 38.2 | - | - | - | - | 56.0 | - | 88.0 | - |
| | Pix2Struct$^*_{large}$ | 1B | 2,048 | 76.6 | 40.0 | - | - | - | - | 58.6 | - | 95.5 | - |
| Token$^V \geq 1k$ | CogAgent | 17B | 6,656 | 81.6 | 44.5 | - | - | - | - | 68.4 | 76.1 | - | - |
| | IXC 2.5 | 7B | $\sim$ 5,118 | 90.9 | 69.9 | **71.2** | - | **53.6** | **85.2** | 82.2 | 78.2 | - | **307.5** |
| | InternVL 2 | 8B | $\sim$ 3,133 | **91.6** | **74.8** | - | - | - | - | **83.3** | **77.4** | | |
| | TokenPacker | 13B | $\sim$ 1,833 | 70.0 | - | - | - | - | - | - | - | - | - |
| | DocOwl 1.5 | 8B | $\sim$ 1,698 | 82.2 | 50.7 | 68.8 | **38.7** | 40.6 | 80.2 | 70.2 | 68.6 | **131.6** | 246.4 |
| | DocPeida | 7B | 1,600 | 47.1 | 15.2 | - | - | - | - | 46.9 | 60.2 | - | - |
| | Monkey | 9B | 1,280 | 66.5 | 36.1 | 40.6 | 32.8 | 25.3 | - | - | 64.3 | 93.2 | - |
| Token$^V < 1k$ | DocOwl | 7B | $\sim$ 841 | 62.2 | 38.2 | 42.6 | 30.3 | 26.9 | 60.2 | 57.4 | 52.6 | 111.9 | 188.8 |
| | UReader | 7B | $\sim$841 | 65.4 | 42.2 | 49.5 | 32.8 | 29.4 | 67.6 | 59.3 | 57.6 | 118.4 | **221.7** |
| | TextMonkey | 9B | 768 | 73.0 | 28.6 | 59.7 | **37.8** | 31.9 | - | 66.9 | 65.9 | - | - |
| | TokenPacker | 13B | $\sim$ 467 | 58.0 | - | - | - | - | - | - | - | - | - |
| | QwenVL | 9B | 256 | 65.1 | 35.4 | - | - | - | - | 65.7 | 63.8 | - | - |
| | Vary | 7B | 256 | 76.3 | - | - | - | - | - | 66.1 | - | - | - |
| | DocOwl2 | 8B | 324 | **80.7** | **46.4** | **66.8** | 37.5 | **36.5** | **78.2** | **70.0** | **66.7** | **131.8** | 217.4 |

Table 9: Comparison with OCR-free Multimodal Large Language Models on single-image document understanding benchmarks. 'FTL(s)' refers to the First Token Latency (seconds)

| Model | Size | DocVQA | | | ChartQA | | | TextVQA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Token$^V$ | FTL(s)↓ | ANLS↑ | Token$^V$ | FTL(s)↓ | ANLS↑ | Token$^V$ | FTL(s)↓ | ANLS↑ |
| InternVL 2 | 8B | $\sim$ 3,198 | 0.94 | 91.6 | $\sim$ 1,827 | 0.56 | 83.3 | $\sim$2,864 | 1.01 | 77.4 |
| IXC 2.5 | 7B | $\sim$7,395 | 3.73 | 90.9 | $\sim$1,971 | 1.05 | 82.2 | $\sim$2,075 | 1.11 | 78.2 |
| DocOwl 1.5 | 8B | $\sim$1,806 | 0.58 | 82.2 | $\sim$1,713 | 0.53 | 70.2 | $\sim$1,664 | 0.56 | 68.6 |
| TextMonkey | 9B | 768 | 0.58 | 73.0 | 768 | 0.51 | 66.9 | 768 | 0.50 | 65.9 |
| DocOwl2 | 8B | 324 | **0.26** | 80.7 | 324 | **0.21** | 70.0 | 324 | **0.23** | 66.7 |

Table 10: Ablation study on DocCompressor with LLaVA-Next-Interleaves. 'FTL(s)' refers to the First Token Latency (seconds). 'R.Acc' refers to Relaxed Accuracy. 'DC' refers to DocCompressor.

| Model | DocVQA | | ChartQA | | MP-DocVQA | | DUDE | | NewsVideoQA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Token$^V$ | ANLS↑ | Token$^V$ | R.Acc↑ | Token$^V$ | ANLS↑ | Token$^V$ | ANLS↑ | Token$^V$ | ANLS↑ |
| LLaVA-Next-Interleaves-7B | $\sim$3,061 | **76.0** | $\sim$1,677 | 69.5 | 729 | 50.0 | 729 | 39.0 | 729 | 62.4 |
| LLaVA-Next-Interleaves-7B w/ DC | 729 | 73.8 | 729 | **73.2** | 729 | **54.2** | 729 | **41.7** | 729 | **64.7** |

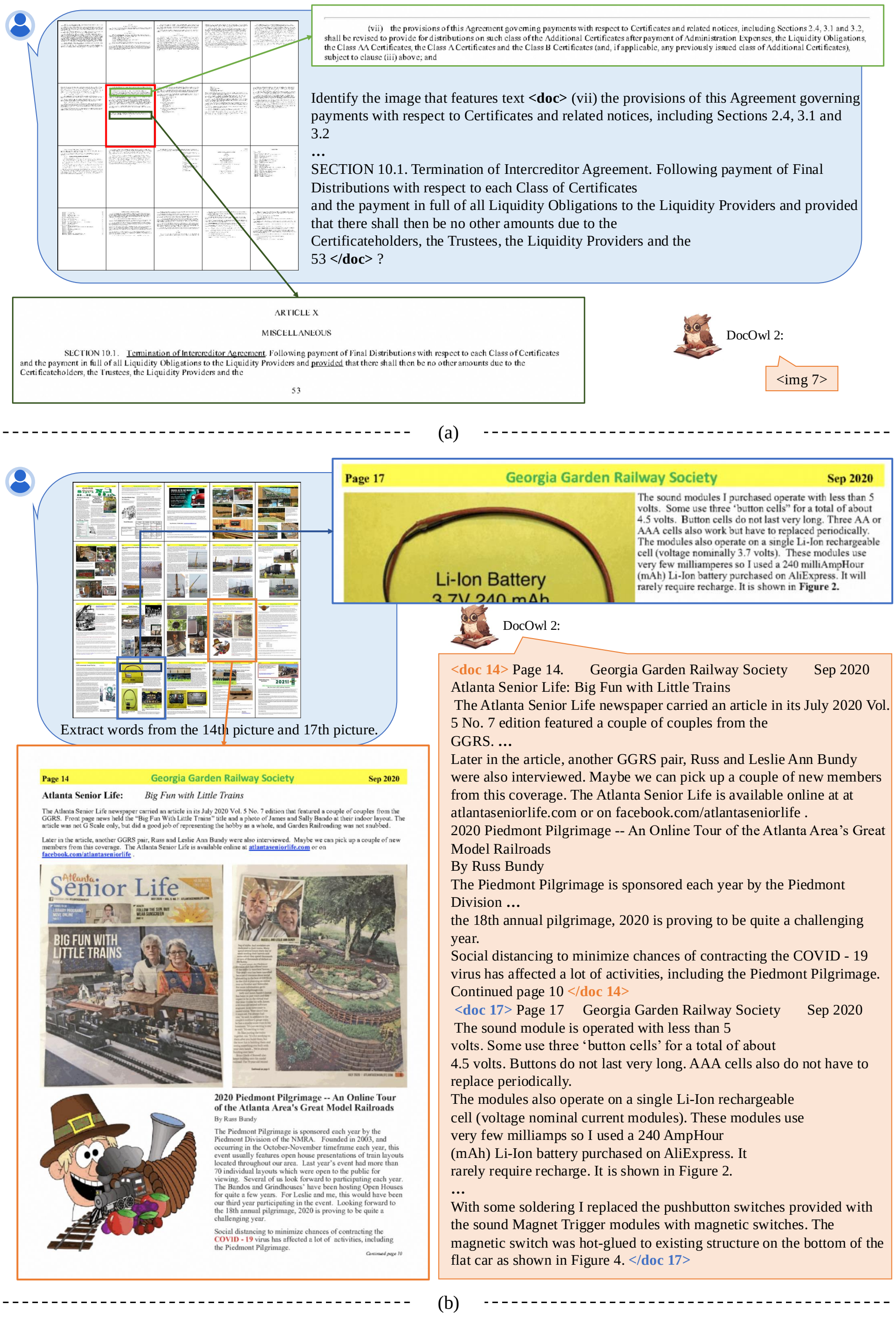



Figure 6: Qualitative results of the Multi-page Text Lookup (a) and Multi-page Text Parsing (b) given by DocOwl2 after the Multi-image Continue Pretraining.

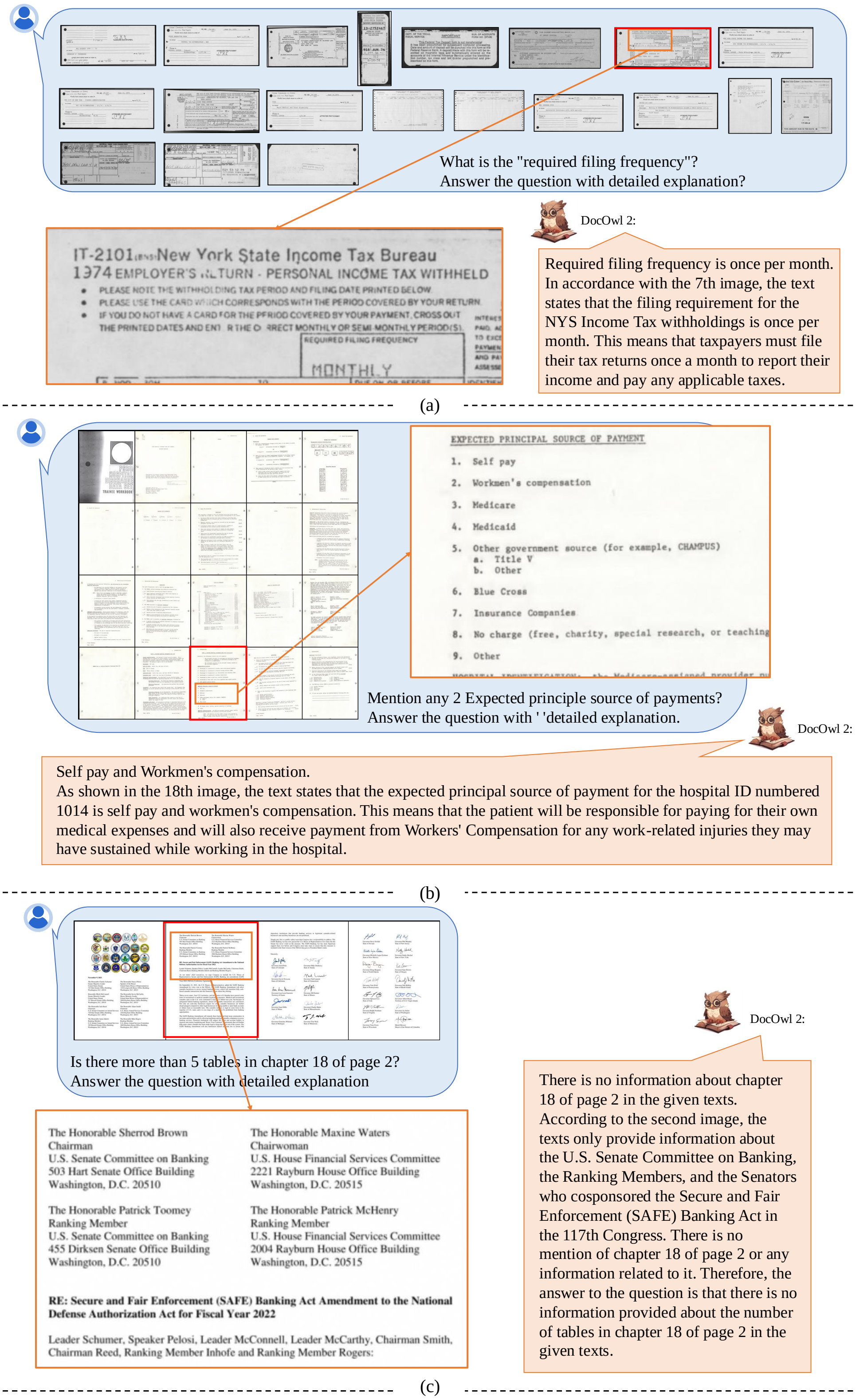



Figure 7: Qualitative results of the Multi-page Question Answering with detailed explanation.


Figure 8: Qualitative results of the Text-rich Video Understanding.