HRVDA: High-Resolution Visual Document Assistant

Chaohu Liu[1,2*§], Kun Yin[3*], Haoyu Cao[3], Xinghua Jiang[3], Xin Li[3]
Yinsong Liu[3], Deqiang Jiang[3], Xing Sun[3], Linli Xu[1,2†]
[1]School of Computer Science and Technology, University of Science and Technology of China
[2]State Key Laboratory of Cognitive Intelligence, [3]Tencent YouTu Lab
liuchaohu@mail.ustc.edu.cn
{zhanyin, rechycao, clarkjiang, fujikoli, jasonysliu, dqiangjiang, winfredsun}@tencent.com
linlixu@ustc.edu.cn

Abstract

*Leveraging vast training data, multimodal large language models (MLLMs) have demonstrated formidable general visual comprehension capabilities and achieved remarkable performance across various tasks. However, their performance in visual document understanding still leaves much room for improvement. This discrepancy is primarily attributed to the fact that visual document understanding is a fine-grained prediction task. In natural scenes, MLLMs typically use low-resolution images, leading to a substantial loss of visual information. Furthermore, general-purpose MLLMs do not excel in handling document-oriented instructions. In this paper, we propose a High-Resolution Visual Document Assistant (**HRVDA**), which bridges the gap between MLLMs and visual document understanding. This model employs a content filtering mechanism and an instruction filtering module to separately filter out the content-agnostic visual tokens and instruction-agnostic visual tokens, thereby achieving efficient model training and inference for high-resolution images. In addition, we construct a document-oriented visual instruction tuning dataset and apply a multi-stage training strategy to enhance the model's document modeling capabilities. Extensive experiments demonstrate that our model achieves state-of-the-art performance across multiple document understanding datasets, while maintaining training efficiency and inference speed comparable to low-resolution models.*

1. Introduction

Large Language Models (LLMs), such as ChatGPT [47], LLaMA [61, 62], have taken a significant stride towards general artificial intelligence. By leveraging massive amounts of data, they have developed powerful reasoning and instruction understanding capabilities. The proliferation of LLMs has also faciliated the development of Multimodal Large Language Models (MLLMs), which can perceive and analyze information from images and other sources [14, 39, 40, 48, 70, 77]. Some existing works have demonstrated that MLLMs exhibit preliminary visual document understanding capabilities, as they can extract and comprehend information from complex documents containing textual and visual elements, such as tables, charts, and graphics [4, 65, 68, 69]. Given their ability to capture the relationships between textual and visual information, employing MLLMs for visual document understanding tasks shows great potential.

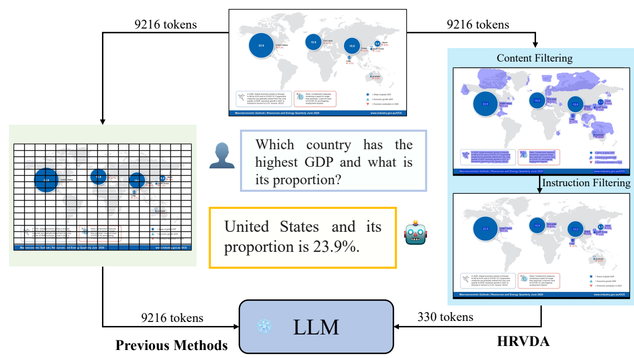

Figure 1. Comparison of the visual processing workflow between HRVDA and previous methods. Previous methods generally employ a low-resolution image encoder to extract features. In contrast, HRVDA utilizes a content filtering mechanism and an instruction filtering module to selectively filter out content-agnostic and instruction-agnostic visual tokens, making high-resolution image processing computationally feasible.

However, the document image processing capabilities of MLLMs are restricted in real-world scenarios, primarily due to two reasons: the limitations posed by low-resolution image inputs and the lack of document-oriented visual in-

* Equal contribution. † Corresponding author. § Work done during an internship at Tencent YouTu Lab.

struction tuning [71].

The restriction of low-resolution image inputs is a prevalent challenge in the multimodal community. Current models usually handle images with relatively low resolutions, typically 224×224 pixels [4, 14, 40]. While this resolution is sufficient for the majority of natural images, it can result in extensive text distortion when it comes to processing document images. As illustrated in Figure 1, clear text in high-resolution images becomes blurred when resized to a lower resolution.

Directly increasing the image resolution generates a large number of visual tokens, which will occupy the limited input capacity of LLMs, and induce considerable training costs and inference latency [17]. Taking CLIP's image encoder [23, 51] as an example, a 1536×1536 image partitioned into 16×16 patches results in 9216 visual tokens, which exceeds the context length of many existing open-source LLMs, such as LLaMA-2 [62] with a context length of 4096. In addition, they exhibit quadratic computational complexity with respect to the length of the patch sequence.

On the other hand, general-purpose MLLMs suffer from a lack of document-oriented visual instruction tuning [40], leading to an incomplete understanding of document images. Unlike ordinary images, document images possess distinct layout and structural information, where the font, style, and color hold significant importance for comprehending the content [45, 56].

To tackle these challenges, we propose a novel multimodal large language model called **HRVDA** (**H**igh-**R**esolution **V**isual **D**ocument **A**ssistant), which employs a **content filtering mechanism** and an **instruction filtering module** designed to filter out content-agnostic visual tokens and instruction-agnostic visual tokens, respectively.

Specifically, content-agnostic visual tokens contribute a significant amount of redundant information, while the regions in document images that contain text, tables, charts, and other document content frequently provide the most valuable information. As shown in Figure 1, the pixels within these regions constitute only a small proportion of the entire image [45]. To reduce the number of blank background tokens, our proposed content filtering mechanism, based on a content detector, can extract key features from document images. Conservatively estimated, this approach filters out approximately 50% of content-agnostic tokens in practice, resulting in a substantial reduction of 30% in training and inference latency without compromising performance.

Meanwhile, instruction-agnostic visual tokens refer to the parts that are not within the instruction attention region. In conventional document understanding tasks, such as information extraction, document-oriented instructions often rely on localized areas to generate answers [30, 49]. Therefore, we design the instruction filtering module to further filter instruction-agnostic visual tokens and significantly reduce the workload of the LLM.

To improve the document understanding capabilities of HDVDA, we construct a document-oriented visual instruction tuning dataset. This dataset covers an extensive array of tasks within the document domain, including information extraction, text recognition, and visual question answering. It also incorporates a variety of scenarios, such as tables, charts, natural images, and webpage screenshots. Furthermore, we employ ChatGPT [47] to generate a diverse collection of instruction templates, thereby strengthening the generalization capabilities of the model.

Our experimental results on multiple document-oriented datasets demonstrate that HRVDA's OCR-free document comprehension capabilities surpass current state-of-the-art MLLMs such as mPLUG-DocOwl [68], UReader [69].

In summary, our main contributionsare as follows:

- We present **HRVDA** (**H**igh-**R**esolution **V**isual **D**ocument **A**ssistant), which, to the best of our knowledge, is the first large multimodal model designed to directly accept high-resolution image inputs.
- We propose a content filtering mechanism and an instruction filtering module to prune visual tokens, which significantly accelerate the model's training and inference, making the processing of high-resolution image inputs computationally feasible.
- We construct an extensive document-oriented visual instruction tuning dataset to enhance the model's document analysis capabilities.
- Experimental results on a series of document-oriented datasets demonstrate that HRVDA achieves state-of-the-art performance.

2. Related Work

2.1. Visual Document Understanding

Visual Document Understanding (VDU) refers to the automated process of analyzing, comprehending, and processing document images [3, 8, 22, 25]. Existing methods can be broadly categorized into two groups, OCR-dependent methods and OCR-free methods.

OCR-dependent methods typically rely on an external OCR interface to extract text content and coordinate information from document images [19, 32, 50, 72]. For instance, the LayoutLM family [29, 66, 67] leverages multimodal pre-training to combine image layout features with textual features. DocFormer [2] undergoes unsupervised pre-training through carefully designed tasks to encourage multimodal interactions. UDOP [60] harmonizes image, text, and layout modalities into a unified and cohesive representation by leveraging the spatial relationships within the document. These methods typically face issues such as increased computational costs and error accumulation [8].

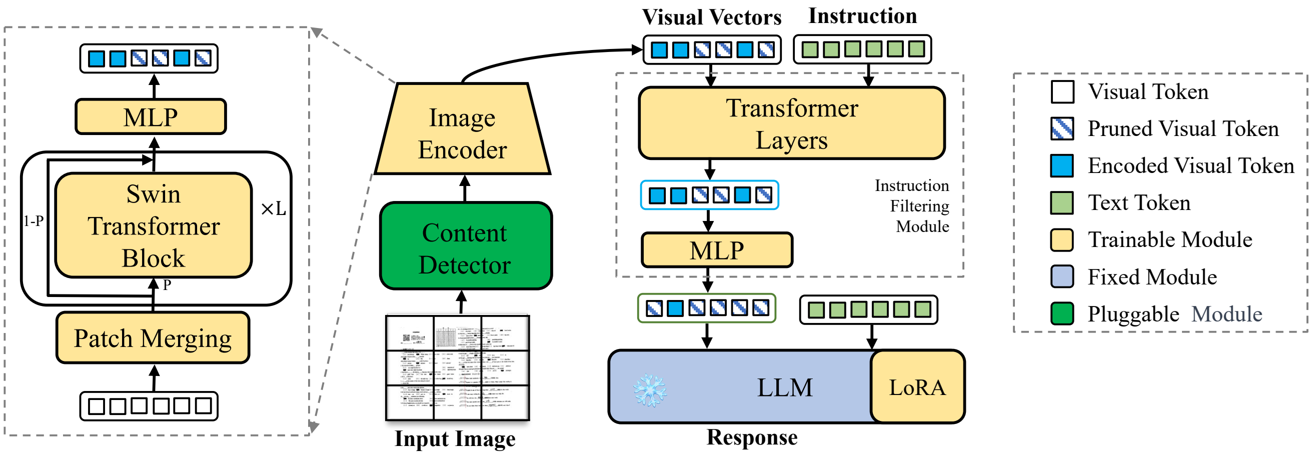

Figure 2. The overall architecture of our proposed HRVDA. After partitioning the document image into visual tokens, a pluggable content detector identifies whether tokens contain document content information, and then a content filtering mechanism is employed to perform token pruning. Encoded visual tokens are then processed through an MLP to maintain consistency with the LLM's embedding space dimensions. The pruned token sequence is fused with the instruction features, further filtering out tokens irrelevant to the instructions. Ultimately, a streamlined set of visual tokens and instructions are fed into the LLM, generating corresponding responses.

OCR-free methods aim to extract structured text directly from images in an end-to-end manner. This approach simplifies the information processing process, speeds up the reasoning and has gained significant attention in the VDU community recently [18, 38]. For example, both Donut [33] and Dessurt [21] utilize Swin Transformer to extract image features, followed by cross-attention operations between decoder models like BART and image features to generate text in an auto-regressive manner. SeRum [9] goes a step further by employing selective region concentration to enhance the precision and speed of generation.

2.2. Multimodal Large Language Models

MLLMs have become a new research focus recently [71]. According to the modality alignment approach, they can be roughly divided into two categories: query-based methods and projection-based methods.

Query-based methods involve utilizing a set of learnable query tokens to extract information through cross-attention mechanisms. Flamingo [1] and BLIP-2 [37] are the first to adopt this approach, which is later inherited by a series of works [13, 20, 70, 73, 77]. However, this method essentially introduces a textual supervisory signal to extract image features and is not suitable for fine-grained prediction tasks. The experimental results are provided in the Appendix A.

Projection-based methods involve directly mapping visual tokens with the LLM's input space [24, 39, 44, 58, 65, 75]. For instance, LLaVA employs a simple linear layer to project image features [40]. LLaMA-Adapter applies a lightweight adapter module to align visual tokens and text tokens [74]. This approach allows the LLM to perceive the entire image, offering a more promising perspective for effective multimodal learning.

2.3. Token Pruning

Token pruning is a technique aimed at reducing model parameters and computational complexity [6, 12, 42, 53]. It achieves model simplification and compression by removing certain weights or feature representations. Numerous methods for pruning vision transformers have been proposed [34]. DynamicViT [52] accelerates model inference by sparsifying less important tokens using lightweight prediction modules. SparseViT [17] efficiently processes high-resolution images through sparse activations, enabling efficient dense prediction tasks. STVit [11] achieves efficient global and local processing in ViTs by removing redundant image tokens and can serve as a backbone for downstream tasks. These pruning techniques are designed for natural images and are not suitable for document images.

3. HRVDA

In this section, we start with the model architecture (in Section 3.1), followed with a detailed explanation of the **Content Filtering Mechanism** (in Section 3.2) and the **Instruction Filtering Module** (in Section 3.3). Finally, we introduce the instruction tuning dataset constructed for document understanding (in Section 3.4) and the training strategy (in Section 3.5).

3.1. Overall Architecture

HRVDA is a large multimodal model designed to address the challenges posed by high-resolution requirements in visual document understanding tasks. As shown in Figure 2, it mainly consists of four modules: a content detector, an image encoder, an instruction filtering module (IFM), and an LLM.

The initial step involves partitioning the original image into a series of patches, which are subsequently converted into a sequence of visual tokens. These tokens are then processed by a content detector to assess the probability of each token containing significant information. Leveraging these probabilities, a content filtering mechanism enables the image encoder to selectively compute visual features and eliminate content-agnostic visual tokens. These encoded visual features are subsequently integrated with the instruction features using a self-attention mechanism within the instruction filtering module. A straightforward 2-layer MLP network is employed to classify these fused features and further exclude instruction-agnostic visual tokens. Ultimately, the highly refined visual tokens are concatenated with the instruction tokens and fed into the LLM for generating the anticipated response. This approach ensures a more efficient and effective representation of the image content, tailored specifically for the task at hand.

3.2. Content Filtering

In conventional Transformer architectures [63], high-resolution images are converted into long token sequences, which poses a substantial demand on computational resources. Moreover, elongated sequences introduce challenges in capturing long-range dependencies.

A potential solution to these challenges lies in the unique properties of document images: they typically consist of extensive areas of blank background, while content-rich regions provide the majority of valuable information [45]. To leverage the sparse content information effectively and efficiently, we propose a content filtering mechanism, primarily involving two modules: the content detector and the image encoder.

Content Detector. A pluggable network is employed to identify whether each token contains important content. For document images, such content includes elements such as text, tables, and charts [45]. The choice of network can be quite diverse. It could be a simple MLP network for token classification, a detection network like DETR [10], or a segmentation network like U-Net [54] applied to reshaped feature maps. In this work, we employ a shallow PSENet [64], which is designed as a segmentation-based detector capable of localizing text instances of any shape. The content detector adopts a high recall rate strategy, ensuring that all visual tokens containing content are preserved.

Image Encoder. A visual backbone network is used to extract image features. In contrast to most MLLMs that utilize ViT [23], we adopt the Swin Transformer [43] as our image encoder, which utilizes a window-based mechanism for self-attention computation, mitigating computational burdens. Moreover, it incorporates a token merge mechanism to prevent the direct loss of information. The Swin Transformer's downsampling of feature maps also contributes to a further reduction of the number of visual tokens.

Given an image $\mathrm{x} \in \mathbb{R}^{H \times W \times C}$, the patch partition module transforms it into a set of visual tokens $\{z_i \mid z_i \in \mathbb{R}^d, 1 \leq i \leq n\}$, where n represents the number of image patches and d is the dimension of the latent vectors of the encoder. The content detector performs a binary classification task on the visual tokens and can obtain the probability $\{p_i \mid p_i \in [0,1], 1 \leq i \leq n\}$ that each patch contains valuable content. Note that the patch partition module employed by the content detector exhibits a structure similar to that of the Swin Transformer, yet they do not share parameters.

As shown on the left side of Figure 2, a skip connection is introduced in each Swin Transformer block to accelerate computation:

$$\begin{aligned} h^{j+1} &= p * F^j(h^j) + (1-p) * h^j \\ h^0 &= z \end{aligned} \tag{1}$$

where F^j represents the operation in the j-th Swin Transformer block, and h^j is the hidden state of the visual tokens.

For a well-trained content detector, we employ a threshold ϵ_c to adjust the probability values in P for tokens containing content:

$$p_i = \begin{cases} 1, & p_i \geqslant \epsilon_c \\ 0, & p_i < \epsilon_c \end{cases} \tag{2}$$

Utilizing these probabilities, if none of the tokens within a window is considered to contain content, the window bypasses the attention computation and is directly passed to the next block, thereby achieving computational acceleration.

It is worth highlighting once again that content-agnostic tokens are not directly removed, making the four merging adjacent patches spatially close. The shifted window partitioning approach [43] in the Swin Transformer enables interactions between different tokens, thereby preserving potentially useful layout information and enhancing modeling capabilities.

The patch merging operation in Swin Transformer consolidates adjacent 2×2 regions into a single new patch, and the probability of the merged patch containing content is set to the maximum value among the probabilities of the 4 original regions:

$$p'_i = \max(p_i, p_{i+1}, p_{i+2}, p_{i+3}) \tag{3}$$

To further preserve global information, the threshold value ϵ is progressively increased from shallow to deeper layers. Preserving more tokens in the shallow layers can reduce the loss of visual information.

Task	Format
DC	Human: What is the category of this image? AI: {cls}
IE	Human: what is the value of the {key}? AI: {value}
VQA	Human: {question} AI: {answer}
OCR	Human: Present all the text in the image. AI: {all text}
VG	Human: Where is the {obj}? AI: {x, y, x + w, y + h}
IC	Human: What is the abstract of the image? AI: {caption}
TR	Human: What is the element in the table? AI: {element}

Table 1. Illustrative examples of instruction tuning templates customized for seven tasks.

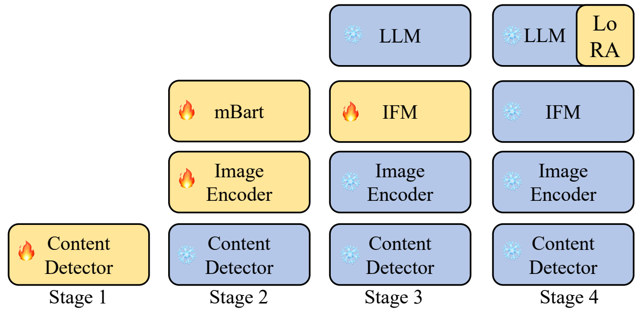

Figure 3. The training pipeline of our HRVDA model.

3.3. Instruction Filtering

Document-oriented instructions are highly specific, typically referring only to particular regions within the image, which necessitates further filtering of visual tokens.

Several existing methods, for instance, the Q-Former module in BLIP-2 [20, 37] and the Visual Abstractor in mPLUG-owl [70], employ learnable queries to extract valuable information. Nevertheless, this approach inadvertently leads to a diminished representation of visual information, making it less suitable for fine-grained prediction tasks. Moreover, the inclusion of query vectors essentially relies on text as a supervisory signal, yet the textual descriptions of images are often insufficient to provide accurate representations. On the other hand, we experimentally discover that for high-resolution images, approximately 500 query vectors are required to maintain performance without significant degradation. This indicates that this approach does not offer a computational advantage in terms of processing speed.

In this study, we utilize a more direct instruction filtering module (IFM) that avoids excessive compression of visual information, thus preserving its integrity.

Formally, the visual vectors obtained from the image encoder and the instruction vectors are concatenated and then fed into the instruction filtering module for further processing. Then, a Transformer layer is employed to facilitate the fusion of these feature vectors:

$$[V', I'] = FFN(SA([V, I])) \tag{4}$$

where SA stands for the self-attention layer, FFN represents the feedforward layer, and V, I denote the visual vectors and instruction vectors, respectively. The fused visual features V' are then sent to a 2-layer MLP for binary classification to filter out visual tokens that are irrelevant to the instructions [42]. Similar to the content detector, the instruction filtering module also adopts a filtering threshold ϵ_i, as in Equation 3, to increase the classification recall rate, ensuring that visual tokens related to instructions are not discarded.

Ultimately, following content-agnostic and instruction-agnostic filtering, the visual token sequences are fed into the LLM.

3.4. Visual Instruction Tuning

In this section, we primarily introduce the task of visual instruction tuning and the data sources.

Tuning Tasks. To enhance HRVDA's generalization in visual document understanding, we organize a wide range of document tasks into an instruction format. In this work, we primarily focus on tasks such as document classification (DC), information extraction (IE), visual question answering (VQA), optical character recognition (OCR), visual grounding (VG), image captioning (IC), and table reconstruction (TR). Table 1 presents some fundamental examples.

To diversify the range of prompts, we first manually craft 10 prompt templates for each task. Subsequently, we employ ChatGPT [47] to generate 50 similar prompts, which are then reviewed by human experts to ensure their alignment with the intended meaning. Additional templates can be found in the Appendix B.1.

Instruction Data Resources. A large number of real-world and synthetic datasets are collected. The real-world datasets used in this study include IIT-CDIP [27], CORD [49], SROIE [30], DocVQA [45], InfographicsVQA [46], DeepForm [7], Kleister Charity [57], WikiTableQuestions [5], TabFact [16], ChartQA [15], TextVQA [56], TextCaps [55], VisualMRC [59], PubTabNet [76], *etc.* Given the limited availability of open source data, in this work a significant amount of data synthesis methods are applied, such as SynthText [26], Synth90K [31] and SynthDoG [33]. Due to space constraints, more details can be seen in the Appendix B.2.

Model	Res.	CORD	SROIE	Doc VQA	Info VQA	Deep Form	KLC	WTQ	Tab Fact	Chart QA	Text VQA	Visual MRC	Text Caps
Donut*	1280	84.1	83.2	67.5	11.6	61.6	30.0	18.8	54.6	41.8	43.5	93.9	74.4
SeRum*	1280	84.9	85.8	71.9	13.5	50.7	31.3	25.5	58.3	47.9	66.3	98.6	101.4
Pix2Struct	1024	-	-	**76.6**	40.0	-	-	-	-	58.6	-	-	-
CogVLM	490	-	-	-	-	-	-	-	-	-	69.7	-	**144.9**
Qwen-VL†	448	-	-	65.1	29.9	2.2	8.9	16.1	52.5	66.3	63.8	76.5	20.25
mPLUG-Doc	224	-	-	62.2	38.2	42.6	30.3	26.9	60.2	57.4	52.6	188.8	111.9
UReader	224	-	-	65.4	42.2	49.5	32.8	29.4	67.6	59.3	57.6	**221.7**	118.4
HRVDA	1536	**89.3**	**91.0**	72.1	**43.5**	**63.2**	**37.5**	**31.2**	**72.3**	**67.6**	**73.3**	211.5	125.3

Table 2. Comparison of HRVDA with OCR-free models across 12 document domain datasets. For consistent comparison, * denotes results obtained after fine-tuning, while † indicates results evaluated based on open-source models. The best results are marked in bold.

Settings	Res.	Encoder	Decoder	All
Qwen-VL	448	1.67	7.8	9.47
HRVDA(0.25, 0.25)	1536	0.92	6.33	7.25
HRVDA(0.25, 0.5)	1536	0.89	4.68	5.57
HRVDA(0.5, 0.25)	1536	0.75	4.05	4.80
HRVDA(0.5, 0.5)	1536	0.76	2.88	3.64

Table 3. Comparison of forward-inference efficiency between HRVDA and Qwen-VL. HRVDA is configured with four sets of filtering thresholds for content and instruction.

3.5. Training Strategies

In order to achieve visual token filtering and enhance the model's document-oriented instruction understanding capabilities, a multi-stage training strategy is adopted in this work as shown in Figure 3.

Stage 1 focuses on training the content detector. We employ external OCR tools and detection networks to obtain the coordinates of various elements, including text, charts, tables, *etc*. These coordinates can be used to provide supervised signals for the PSENet, determining whether each visual token contains content or not. **Stage 2** concentrates on the pretraining of the image encoder. Our encoder is integrated with m-BART [41] via cross-attention to perform the task of recognizing all text within the images [33]. **Stage 3** involves the training of the instruction filtering module. For data with fixed layouts, a high filtering threshold is used. Conversely, we utilize a low filtering threshold for data characterized by variable layouts. **Stage 4** entails implementing low-rank adaptation techniques to preserve the general conversational capabilities of the LLM [28]. Additional training details can be found in the Appendix C.

4. Experiments

In this section, we conduct experiments on numerous publicly available document-oriented datasets to validate the effectiveness of our proposed HRVDA model.

4.1. Tasks and Datasets

In visual document understanding, information extraction and text-oriented visual question answering are challenging tasks, which also have widespread applications in practice.

Information Extraction involves extracting structured key-value pair data from documents. In this study, we use the two most commonly used datasets for evaluation, CORD [49] and SROIE [30]. They are all scanned receipt images and have good image quality. The F1 score is reported, which is the weighted harmonic mean of Precision and Recall.

Text-oriented Visual Question Answering is a highly generalizable task, capable of addressing various problems through appropriate prompts. We evaluate HRVDA on a wide range of publicly available datasets, including DocVQA [45], InfoVQA [46], TextVQA [56], ChartQA [15], DeepForm [7], KLC [57], WTQ [5], TableFact [16], VisualMRC [59], and TextCaps [55]. Different metrics, including ANLS, CIDEr, Accuracy, and F1 Score are reported in accordance with the methodologies employed in previous works. A detailed description can be found in the Appendix B.2.

4.2. Implementation Details

Model Architecture. Our HRVDA model employs Swin-L [43] as the image encoder. Its layer and window sizes are set to 2, 2, 18, 2, and 10, respectively, with a patch size of 4×4. Additionally, the image resolution is set to 1536×1536. In this study, we conduct experiments based on LLaMA-2-7B [62], which has a context length of 4096.

Training Details. We employ the Adam optimizer for each stage of training, with an initial learning rate of 1e-4. The learning rate schedule uses a linear warmup during the first 20% of steps. For LoRA, we set the rank to 8. Unless otherwise specified, the detection thresholds for content filtering in the Swin Transformer are set to $\epsilon_c = [0.25, 0.25, 0.5, 0.5]$ in 4 stages, while the threshold for instruction filtering is set to $\epsilon_i = 0.5$. The batch size is set at 128. All training is conducted on 128 Tesla V100 GPUs for 10 epochs.

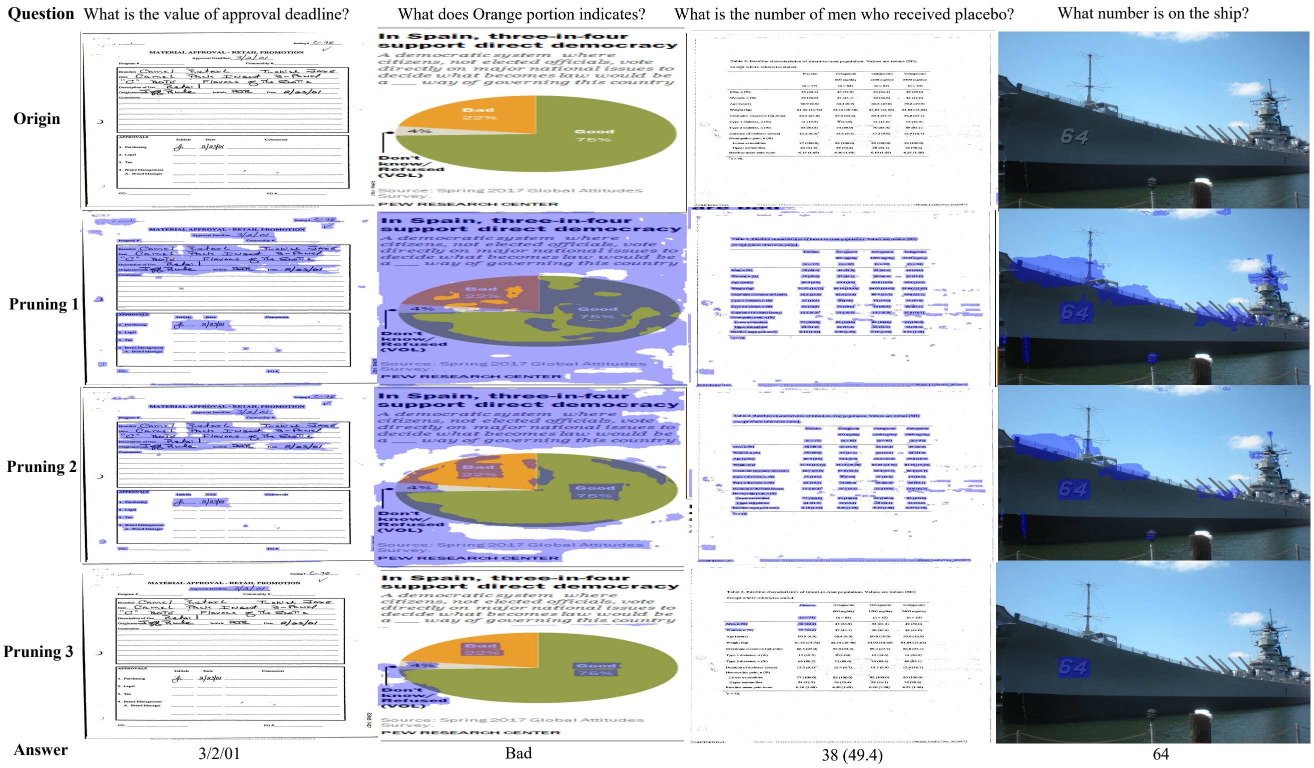

Figure 4. Visualization of the visual token filtering. The first row displays the original images, while the following three rows show the effects of visual token filtering. The Pruning 1 and Pruning 2 occur in the first two stages and the last two stages of the Swin Transformer, respectively, while Pruning 3 takes place in the instruction filtering module.

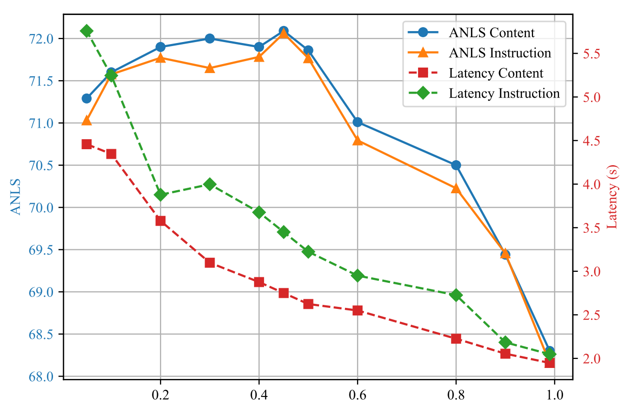

Figure 5. The impact of filtering thresholds on the DocVQA dataset. Best viewed in color.

4.3. Comparisons with Previous Approaches

We conduct a comparative analysis of HRVDA against OCR-free models, including Donut [33], SeRum [9], Pix2Struct [35], Qwen-VL [4], mPLUG-Doc [68], and UReader [69], utilizing 12 publicly available datasets for evaluation.

These models can be broadly categorized into two classes: encoder-decoder models and MLLMs. The first class utilizes a cross-attention mechanism [63] to fuse image and text, resulting in computational efficiency for high-resolution image inputs while simultaneously requiring task-specific fine-tuning. The second class leverages LLMs, offering exceptional understanding capabilities, but often unable to directly process high-resolution inputs.

As demonstrated in Table 2, HRVDA achieves the best results across the 9 datasets. In information extraction tasks, our model significantly surpasses current state-of-the-art performance, owing to our robust visual pretraining (Stage 2). In visual question answering tasks, understanding the question becomes crucial, particularly in datasets with a high prevalence of elements from natural scene [56]. The semantic analysis capabilities of the decoder in the first category are limited, which prevents them from achieving optimal performance. Previous MLLMs are constrained by the visual information distortion caused by low-resolution image input, which also prevents them from achieving desirable results. Consequently, our HRVDA model directly processes high-resolution image inputs, minimizing the loss of visual information and thereby delivering substantial performance enhancements.

In terms of efficiency evaluation, we use Qwen-VL as our baseline and evaluate the forward-inference latency on a Tesla V100 GPU. The results reveal that HRVDA's speed is significantly faster than Qwen-VL's across various filtering thresholds, as illustrated in Table 3. Remarkably, when

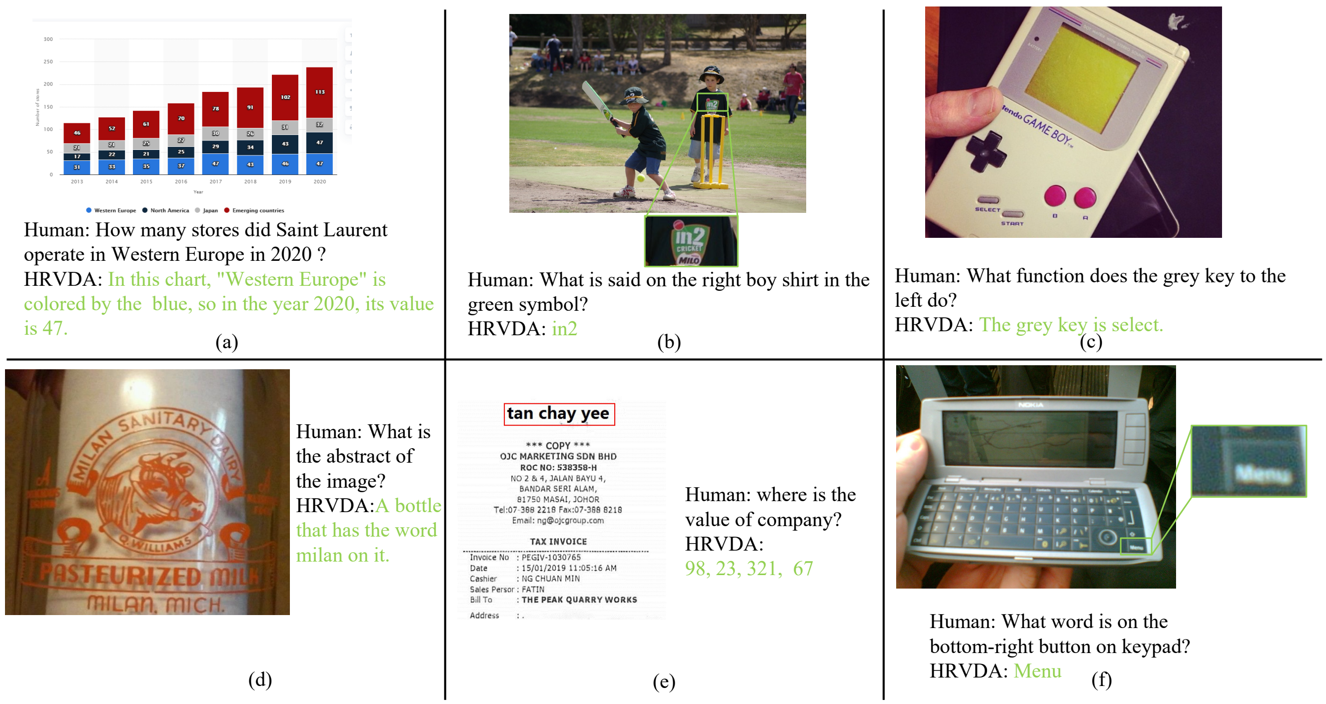

Figure 6. Qualitative examples generated by HRVDA. For better clarity, key regions are magnified and cropped.

both thresholds are set to 0.5, HRVDA reduces the runtime by 61%. However, due to the constraints of GPU memory usage, we do not further increase the resolution.

4.4. Ablation Study

In this section, we separately explore the impact of filtering thresholds in the visual filtering mechanism and instruction filtering module.

Figure 4 showcases several examples of token pruning. It can be observed that for text-dense images, the proportion of filtered pixels is considerably high. In contrast, for images containing charts and natural elements, the filtering ratio is lower, as more visual semantic information is required for these types of images. On the other hand, we quantitatively evaluate the impact of filtering thresholds in the content filtering mechanism and instruction filtering module on prediction accuracy and inference latency, as shown in Figure 5. As the threshold increases, the accuracy of the prediction gradually improves, reaching its peak at 50% and then experiencing a decline. The inference latency decreases almost linearly with the filtering threshold. These results indicate that appropriate token pruning not only accelerates computation but also improves performance, as removing redundant information can reduce the difficulty for the model to extract key information.

4.5. Qualitative Analyzes

As shown in Figure 6, HRVDA can recognize text in specific areas based on location hints. This is extremely useful in practical applications, as people often describe vague locations to obtain information. HRVDA also successfully identifies the highly blurred text *Menu*, which may be due to the influence of visual semantic cues. Utilizing comprehensive document-oriented visual instruction tuning, HRVDA exhibits outstanding capabilities in following document instructions. More cases can be found in the Appendix D.

5. Conclusion

In this work, we propose a new OCR-free multimodal large language model, HRVDA, which can directly accept high-resolution image inputs and is suitable for fine-grained prediction tasks. To the best of our knowledge, HRVDA is the first MLLM to utilize the Swin Transformer as an encoder. Additionally, we employ a content filtering mechanism and an instruction filtering module to alleviate the computational challenges brought about by high-resolution inputs. Experimental results demonstrate that our HRVDA model achieves state-of-the-art results on a series of public datasets, while also exhibiting significantly faster speeds compared to previous MLLMs. In the future, we will continue to investigate high-resolution challenges.

Acknowlegement

This research was supported by the National Key Research and Development Program of China (Grant No. 2022YFB3103100), the National Natural Science Foundation of China (Grant No. 62276245).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 3

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 973–983. IEEE, 2021. 2

[3] Haoli Bai, Zhiguang Liu, Xiaojun Meng, Wentao Li, Shuang Liu, Yifeng Luo, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, Xin Jiang, and Qun Liu. Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13386–13401. Association for Computational Linguistics, 2023. 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2, 7

[5] Jonathan Berant, Daniel Deutch, Amir Globerson, Tova Milo, and Tomer Wolfson. Explaining queries over web tables to non-experts. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 1570–1573. IEEE, 2019. 5, 6, 2

[6] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[7] Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. DUE: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5, 6, 2

[8] Haoyu Cao, Xin Li, Jiefeng Ma, Deqiang Jiang, Antai Guo, Yiqing Hu, Hao Liu, Yinsong Liu, and Bo Ren. Query-driven generative network for document information extraction in the wild. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4261–4271. ACM, 2022. 2

[9] Haoyu Cao, Changcun Bao, Chaohu Liu, Huang Chen, Kun Yin, Hao Liu, Yinsong Liu, Deqiang Jiang, and Xing Sun. Attention where it matters: Rethinking visual document understanding with selective region concentration. *CoRR*, abs/2309.01131, 2023. 3, 7

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229. Springer, 2020. 4

[11] Shuning Chang, Pichao Wang, Ming Lin, Fan Wang, David Junhao Zhang, Rong Jin, and Mike Zheng Shou. Making vision transformers efficient from A token sparsification view. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6195–6205. IEEE, 2023. 3

[12] Shuning Chang, Pichao Wang, Ming Lin, Fan Wang, David Junhao Zhang, Rong Jin, and Mike Zheng Shou. Making vision transformers efficient from A token sparsification view. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6195–6205. IEEE, 2023. 3

[13] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-LLM: bootstrapping advanced large language models by treating multi-modalities as foreign languages. *CoRR*, abs/2305.04160, 2023. 3

[14] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 2

[15] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 5, 6, 2

[16] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 5, 6, 2

[17] Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2061–2070. IEEE, 2023. 2, 3

[18] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. Pali-3 vision language models: Smaller, faster, stronger. *CoRR*, abs/2310.09199, 2023. 3

[19] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document AI: benchmarks, models and applications. *CoRR*, abs/2111.08609, 2021. 2

[20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023. 3, 5

[21] Brian L. Davis, Bryan S. Morse, Brian L. Price, Chris Tensmeyer, Curtis Wigington, and Vlad I. Morariu. End-to-end document recognition and understanding with dessurt. In *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, pages 280–296. Springer, 2022. 3

[22] Yihao Ding, Zhe Huang, Runlin Wang, Yanhang Zhang, Xianru Chen, Yuzhong Ma, Hyunsuk Chung, and Soyeon Caren Han. V-doc : Visual questions answers with documents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 21460–21466. IEEE, 2022. 2

[23] 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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 4

[24] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter V2: parameter-efficient visual instruction model. *CoRR*, abs/2304.15010, 2023. 3

[25] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4573–4582. IEEE, 2022. 2

[26] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2315–2324. IEEE Computer Society, 2016. 5

[27] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 991–995. IEEE Computer Society, 2015. 5

[28] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 6

[29] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4083–4091. ACM, 2022. 2

[30] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. ICDAR2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1516–1520. IEEE, 2019. 2, 5, 6, 1

[31] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *CoRR*, abs/1406.2227, 2014. 5

[32] Anoop R. Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2d documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4459–4469. Association for Computational Linguistics, 2018. 2

[33] 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 *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, pages 498–517. Springer, 2022. 3, 5, 6, 7, 1

[34] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 784–794. ACM, 2022. 3

[35] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 18893–18912. PMLR, 2023. 7

[36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 12888–12900. PMLR, 2022. 1

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 3, 5

[38] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *CoRR*, abs/2311.06607, 2023. 3

[39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1, 3

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023. 1, 2, 3

[41] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8: 726–742, 2020. 6, 1

[42] Yifei Liu, Mathias Gehrig, Nico Messikommer, Marco Cannici, and Davide Scaramuzza. Revisiting token pruning for object detection and instance segmentation. *CoRR*, abs/2306.07050, 2023. 3, 5

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 4, 6

[44] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *CoRR*, abs/2305.15023, 2023. 3

[45] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE, 2021. 2, 4, 5, 6

[46] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE, 2022. 5, 6, 2

[47] OpenAI. ChatGPT. `https://openai.com/blog/chatgpt/`, 2023. 1, 2, 5

[48] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 1

[49] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing. 2019. 2, 5, 6, 1

[50] Rafal Powalski, Lukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Palka. Going full-tilt boogie on document understanding with text-image-layout transformer. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, pages 732–747. Springer, 2021. 2

[51] 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 2

[52] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13937–13949, 2021. 3

[53] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10843–10852. IEEE, 2022. 3

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015. 4

[55] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: A dataset for image captioning with reading comprehension. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, pages 742–758. Springer, 2020. 5, 6, 2

[56] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE, 2019. 2, 5, 6, 7

[57] Tomasz Stanislawek, Filip Gralinski, Anna Wróblewska, Dawid Lipinski, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemyslaw Biecek. Kleister: Key information extraction datasets involving long documents with complex layouts. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, pages 564–579. Springer, 2021. 5, 6, 2

[58] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355, 2023. 3

[59] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13878–13888. AAAI Press, 2021. 5, 6, 2

[60] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19254–19264. IEEE, 2023. 2

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 1

[62] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. 1, 2, 6

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4, 7

[64] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9336–9345. Computer Vision Foundation / IEEE, 2019. 4

[65] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 1, 3

[66] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1192–1200. ACM, 2020. 2

[67] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2579–2591. Association for Computational Linguistics, 2021. 2

[68] 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. mplug-docowl: Modularized multimodal large language model for document understanding. *CoRR*, abs/2307.02499, 2023. 1, 2, 7

[69] 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 Alex Lin, and Fei Huang. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *CoRR*, abs/2310.05126, 2023. 1, 2, 7

[70] 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. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023. 1, 3, 5

[71] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. 2, 3

[72] Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. PICK: processing key information extraction from documents using improved graph learning-convolutional networks. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 4363–4370. IEEE, 2020. 2

[73] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *CoRR*, abs/2306.02858, 2023. 3

[74] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *CoRR*, abs/2303.16199, 2023. 3

[75] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: visual instruction tuning for medical visual question answering. *CoRR*, abs/2305.10415, 2023. 3

[76] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno-Yepes. Image-based table recognition: Data, model, and evaluation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, pages 564–580. Springer, 2020. 5

[77] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023. 1, 3