VisFocus: Prompt-Guided Vision Encoders for OCR-Free Dense Document Understanding

Ofir Abramovich[1]*, Niv Nayman[2]†, Sharon Fogel*, Inbal Lavi*, Ron Litman[2], Shahar Tsiper[2], Royee Tichauer[2], Srikar Appalaraju[2], Shai Mazor[2], and R. Manmatha[2]

[1] Reichman University, Israel [2] AWS AI Labs

Abstract. In recent years, notable advancements have been made in the domain of visual document understanding, with the prevailing architecture comprising a cascade of vision and language models. The text component can either be extracted explicitly with the use of external OCR models in OCR-based approaches, or alternatively, the vision model can be endowed with reading capabilities in OCR-free approaches. Typically, the queries to the model are input exclusively to the language component, necessitating the visual features to encompass the entire document. In this paper, we present *VisFocus*, an OCR-free method designed to better exploit the vision encoder's capacity by coupling it directly with the language prompt. To do so, we replace the down-sampling layers with layers that receive the input prompt and allow highlighting relevant parts of the document, while disregarding others. We pair the architecture enhancements with a novel pre-training task, using language masking on a snippet of the document text fed to the visual encoder in place of the prompt, to empower the model with focusing capabilities. Consequently, VisFocus learns to allocate its attention to text patches pertinent to the provided prompt. Our experiments demonstrate that this prompt-guided visual encoding approach significantly improves performance, achieving state-of-the-art results on various benchmarks.

Keywords: Document Understanding · OCR-free Models

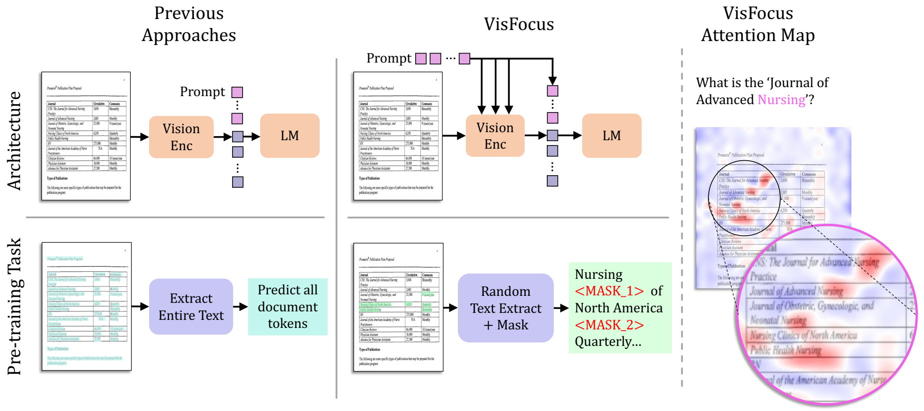
1 Introduction

Visual Document Understanding (VDU) aims to extract meaningful information from digitized documents, such as PDFs or images, extending beyond the scope of Optical Character Recognition (OCR). This field encompasses various tasks, including DocVQA [32], ChartQA [30], InfoVQA [31], Key-Value Identification in forms [17], Entity Extraction [38] and Document Classification [14,18].

*Work done during an internship\employment at Amazon

† Corresponding author: nivnay@amazon.com

Fig. 1: VisFocus' key contributions. The left side of the figure illustrates how VisFocus enables the vision model to better align visual features to the input prompt; Unlike previous approaches, VisFocus inputs the prompt not only to the language model, but to the vision encoder as well (top left vs top middle). In addition, a novel pre-training task utilizes the enabled interactions with the prompt to focus the model on specific text patches (bottom middle) instead of the entire text (bottom left). The right side of the figure shows the resulting attention map from VisFocus illustrating how the model focuses on a specific word taken from the query ('Nursing').


Previous Approaches
VisFocus
VisFocus Attention Map
Architecture
Pre-training Task
Prompt
Vision Enc
LM
Prompt
Vision Enc
LM
What is the 'Journal of Advanced Nursing'?
Extract Entire Text
Predict all document tokens
Random Text Extract + Mask
Nursing <MASK_1> of North America <MASK_2> Quarterly...

Traditional VDU approaches rely on OCR to extract textual information from the document [2, 3, 19, 35, 41, 42, 46, 48, 49]. These Vision-Language (VL) approaches then take OCR text and spatial features as well as visual tokens of the document as input to generate predictions. However, running OCR at training and inference as a pre-processing step adds additional latency and computational costs [10, 22]. In addition, errors originated in the OCR step might propagate to the VL model and deteriorate its performance [20, 40].

The OCR-free [10, 22, 23] approach emerged as an alternative way for VDU. Here, the document image is directly fed as input to the vision-language model, bypassing the need for explicit OCR text extraction. The architecture usually consists of a visual encoder, followed by a language model [6, 10, 22, 23, 50–52] which receives as input the visual representation as well as the input query. The model is expected to internally first learn to read, and then perform the downstream task. To avoid the need for OCR input, an extensive pre-training stage is performed to endow the vision model with reading capabilities [10, 22, 23].

In most OCR-free VDU prior art [10, 22], the user query is used as an input to the language model alone, as illustrated on the top left of Fig. 1. Specifically, since the visual features are processed independently of the input language query we posit that the visual features could be sub-optimal, containing information irrelevant to the user query. This misalignment is particularly critical for dense documents, that contain a large amount of text. For such documents, reading the text properly requires high-resolution input images, containing many pixels of blank areas, figures and text irrelevant to the user query. Intuitively, these can draw a large portion of the encoded visual tokens, while missing sufficient focus on the desired query.

We suggest a new approach for OCR-free VDU, VisFocus, to generate prompt-aware visual features. This is achieved by (1) incorporating the user prompt directly in the vision-encoder architecture (top middle of Fig. 1) and (2) proposing a complimentary pre-training scheme (bottom middle of Fig. 1), that through this coupling, enables the prompt to focus the model on the relevant text in the document (right part of Fig. 1). Our approach is inspired by the selective scanning method one might employ when searching for an answer within a document: rather than meticulously reading through every word, attention is focused on identifying keywords pertinent to the question. Once these keywords are identified, closer scrutiny is applied to the surrounding text to extract the desired answer. Similarly, in VisFocus, the language prompt assigns more weight to relevant visual features by repeatedly merging visual patches of the input document through designated cross-attention mechanism [45] with the input prompt. We term the newly introduced layers ***V**ision-**L**anguage **M**erging **A**ttention* (*ViLMA*) layers. We empirically show that those carefully located ViLMA layers lead to better alignment of visual and language information, enabling the model to focus on contextually relevant information associated with the language prompt, as illustrated in Fig. 1 (right).

Once the interactions between the textual prompt and visual features are enabled within the model architecture, a complimentary pre-training stage is devised. This newly introduced pre-training task, ***L**ocalized **M**asked **P**rompt **M**odeling* (*LMPM*), leverages these interactions for guiding the model to search for semantically-related text to the prompt rather than reading the entire document. This task is illustrated on the bottom middle of Fig. 1. The underlying concept is to enable the model to develop hierarchical understanding of the document, initially learning general reading skills [10, 22] and subsequently refining its ability to focus on specific parts of the text during the second stage.

By carefully combining the ViLMA layers and the the LMPM task, the visual encoder learns to focus its attention on the most relevant patches of the input document. Our empirical analysis shows the contributions of the suggested components (see Sec. 4.3); the LMPM pre-training stage, the ViLMa layers. Thus, exhibiting a symbiosis between those architectural and pre-training enhancements. To summarize our key contributions:

1. We propose novel patch-merging layers, termed ViLMA , which imbue the visual encoding process with prompt awareness, resulting in improved alignment between vision and language for VDU tasks.
2. A new pre-training task tailored for OCR-free VDU, named LMPM, is introduced. This task encourages the visual encoder to extract visual features relevant for the specific prompt, enhancing the model's ability to discern relevant information.
3. Through extensive experimentation, we showcase the synergistic impact of combining these architecture enhancements with the introduced pre-training task. This leads to state-of-the-art performance on multiple benchmarks compared to previous similarly sized OCR-Free methods.

2 Related Work

Document understanding approaches have been widely explored in recent years. In this domain, two primary approaches have emerged: (a) OCR-based approaches, which rely on document OCR to interpret document images, and (b) OCR-free approaches, which bypass OCR to directly solve high-level tasks.

OCR-based Initially, research in VDU predominantly relied on OCR-based approaches, where a general natural language model is employed alongside spatial features extracted from 2D document images, in conjunction with pre-extracted OCR text [2, 3, 19, 34, 35, 41, 42, 46, 48, 49]. In recent years, advancements in this field have been notable. Modern methodologies, such as DocFormer [2,3] and LayoutLM [19,48,49] leverage transformer-based architectures. These models are specifically designed to integrate spatial features, text tokens, and their corresponding bounding boxes into rich representations, enabling more effective document understanding. Additionally, frameworks like UDOP [42] aim at establishing aligned representations of spatial and textual embeddings, which are then fed into a unified encoder. This strong reliance on OCR presents several drawbacks: (1) it relies on off-the-shelf tools, making it susceptible to their errors; (2) it increases the complexity of models and computational overhead (3) processing the entire extracted text can result in unnecessary computations, especially in cases where only specific regions or aspects of the document are relevant to the task at hand. This paved the way for OCR-free approaches.

OCR-free Donut [22] and Dessurt [10] were pioneering works in the realm of OCR-free VDU, introducing models that exclusively process document images without reliance on OCR. These works have shown the significance of equipping models with reading capabilities during pre-training to enhance downstream task performance. Subsequently, Pix2Struct [23] demonstrated performance gains by training larger models with expanded datasets and incorporating additional pre-training tasks. Notably, Pix2Struct opts for rendering the prompt onto the input image, integrating it visually rather than inputting it directly to the language model. This makes the visual features prompt-dependant; however, the use of rendering limits the semantic usefulness of the input prompt. In contrast, VisFocus leverages the prompt in a manner that facilitates semantic understanding, enabling the visual features to prioritize relevant information more effectively. A concurrent work to ours [15] injects the user prompt to arbitrary self-attention layers [45] of a ViT [13] encoder to promote the alignment of visual and lingual features of VL models. While excelling on scene-text images of typically few words, it lacks the crucial complementary pre-training task for reading text segments relevant to the prompt and thus reports low performance for VDU.

A notable category of OCR-free methods is Large Vision-Language Models (LVLMs), including Qwen-VL [7], PaLi-3 [9], MPlug-DocOwl [50], ScreenAI [4] and others [1, 24, 26, 52, 54]. Their main theme is scaling up both the vision encoder, e.g. to ViT-L, ViT-H or ViT-G [12,53], and the LM to Large Language

Models (LLMs), e.g. [5, 43, 44]. Those two components are often connected by advanced alignment modules [26, 54], such as Q-Former [24]. Altogether, these components accumulate to models with billions of parameters, necessitating a significant amount of memory, computational resources, and training data to achieve their superior performance in VDU tasks.

3 VisFocus

We suggest a new OCR-free document understanding method called VisFocus. Our approach revolves around the need to enable interaction between the visual features and the language prompt. To do so, new layers named ViLMA are incorporated into the vision encoder architecture as described in Sec. 3.1. During these interactions, a specifically designed pre-training stage (LMPM) guides the vision encoder to concentrate on the pertinent text patches within the document image in relation to the prompt(Sec. 3.2).

3.1 Architecture Enhancements Enable Early Prompt Interactions

OCR-free architectures are typically composed of two main components: a visual encoder $\mathcal{M}_V$ responsible for encoding an input document image X into visual features $\hat{Z}$:

$$\hat{Z} = \mathcal{M}_{V}\left(X\right) \tag{1}$$

and a language model $\mathcal{M}_L$, that receives both the encoded image and the user prompt $\mathbf{p}$ as an input to produce the final prediction $\hat{\mathcal{Y}}$:

$$\hat{\mathcal{Y}} = \mathcal{M}_{L}\left(\mathbf{p}, \hat{Z}\right) \tag{2}$$

Our objective is to improve the performance over document understanding tasks by introducing the prompt sequence earlier in the model. To this end, instead of having a visual representation of the input document independent of the prompt, VisFocus's encoder $\mathcal{M}_V^p$ encodes the document image with respect to the language prompt (a question, instruction, etc.) to produce prompt-aware features $\hat{Z}_{\mathbf{p}}$:

$$\hat{Z}_{\mathbf{p}} = \mathcal{M}_{V}^{p}\left(\mathbf{p}, X\right) \tag{3}$$

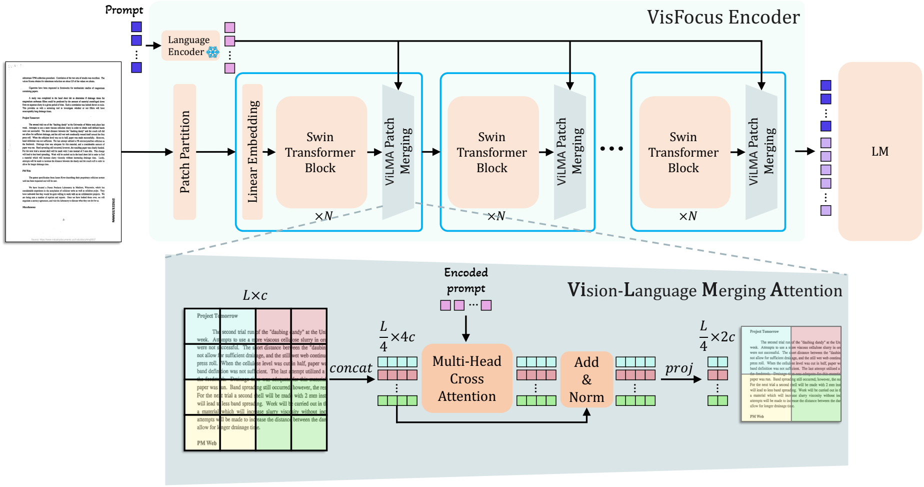
When integrated with our proposed pre-training scheme which we present in Sec. 3.2, this enables the encoder to focus on more relevant text patches of the image with respect to the prompt, as demonstrated at the right side of Fig. 1.

In both previous approaches and in ours, the prompt is also used as an input to the language model, such that the overall model $\mathcal{M}$ formulation reads:

$$\mathcal{M}\left(\mathbf{p}, X\right) = \mathcal{M}_{L}(\mathbf{p}, \hat{Z}_{\mathbf{p}}) = \mathcal{M}_{L}\left(\mathbf{p}, \mathcal{M}_{V}^{p}\left(\mathbf{p}, X\right)\right) \tag{4}$$

Next we specify the architectural enhancements in $\mathcal{M}_V^p$, to enable interaction between visual feature maps and the prompt $\mathbf{p}$.

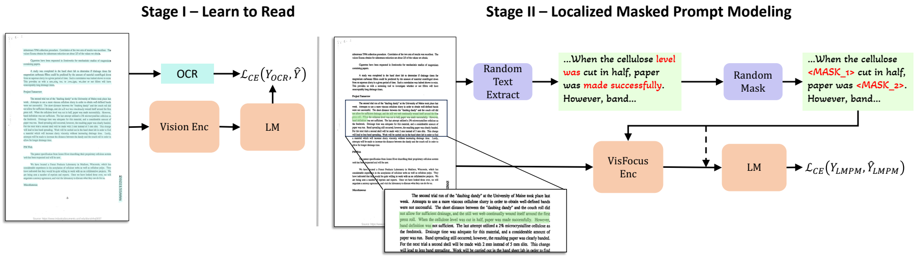
Fig. 2: An overview of the VisFocus architecture. The encoded prompt serves as an input for every ViLMA layer, at the end of each encoding stage (top). The goal of the ViLMA layers is to provide the encoder with prompt guidance during the down-sampling process. The encoded prompt is input through a cross attention layer before down-sampling (bottom).



Vision-Language Merging Attention The vision encoder $\mathcal{M}_V$ of VisFocus is based on the Swin transformer [27], chosen for its hierarchical architecture designed to capture both local and global information effectively. The model merges neighbouring patches via patch merging layers, and thus aggregates information into larger more abstract representations at higher scales. To promote the model's attention towards patches relevant to the input prompt, we chose to intervene at the patch merging layers and modify them accordingly, ensuring that the captured information is contingent on the prompt. We term the newly introduced patch merging mechanism ViLMA, which stands for VIsion-Language Merging Attention. The upper part of Fig. 2 illustrates where the ViLMA layers are incorporated in SwinV2 transformer instead of the original patch merging layers, and the mechanism of a single ViLMA layer is shown at the bottom.

Similarly to the original Swin patch-merging layers, ViLMA concatenates the features of each group of 2×2 neighbouring patches, creating a feature map of size $L/4 \times 4c$ from the original feature map of size $L \times c$, where L is the spatial dimension and c is the corresponding number of channels. Then, a linear layer is applied to aggregate the spatial information into higher-level features, concurrently decreasing the feature count by a factor of 2, yielding a feature map with dimensions $L/4 \times 2c$. To ensure that the feature reduction aligns with the prompt, we introduce an interaction layer between the visual features and the prompt. This cross-attention layer is incorporated just before the projection layer, enabling down-sampling to be conducted relative to the prompt. Following [16, 47] rather than utilizing the original prompt $\mathbf{p}$ to guide the encoding of visual features, we employ a frozen language encoder to generate

Fig. 3: Training Scheme. Previous methods only trained the model to read by predicting the OCR of the document (Stage I). We suggest an addition Localized Masked Prompt Modeling (Stage II) step to train the model to focus on a specific area of text inside the document.


Stage I – Learn to Read
OCR
$\mathcal{L}_{CE}(Y_{OCR}, \hat{Y})$
Vision Enc
LM
Stage II – Localized Masked Prompt Modeling
Random Text Extract
...When the cellulose level was cut in half, paper was made successfully. However, band...
Random Mask
...When the cellulose <MASK_1> cut in half, paper was <MASK_2>. However, band...
VisFocus Enc
LM
$\mathcal{L}_{CE}(Y_{LMPM}, \hat{Y}_{LMPM})$

a context-aware representation of the prompt: emb($\mathbf{p}$). Subsequently, the visual feature map, $\hat{F}$, is passed through a Multi-Head Cross Attention (MHCA) layer [45], together with the prompt encoding. This is followed by normalization and additive layers. Thus, for every feature map:

$$\tilde{F} = \hat{F} + \text{Norm}\left(\text{MHCA}\left(\hat{F}, \text{emb}(\mathbf{p})\right)\right) \tag{5}$$

where in the cross attention layer, the visual feature map $\hat{F}$ is used as the query and the prompt embeddings emb($\mathbf{p}$) are used as both the keys and values.

3.2 Pre-training Scheme for Prompt-Aware Focusing

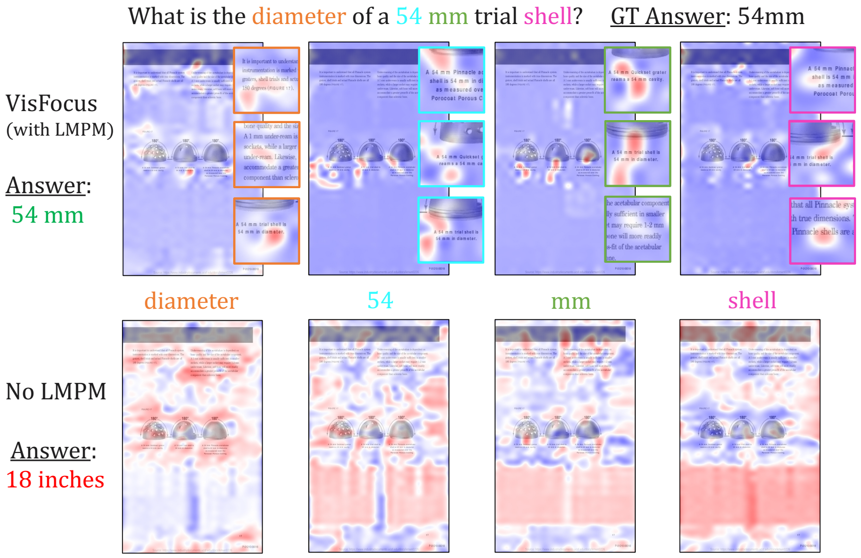
While incorporating ViLMA facilitates the model's interaction with the user prompt during downsampling, it does not explicitly guide the vision encoder towards focusing on the most relevant text patches. To address this, we introduce a new pre-training task called *LMPM* (Localized Masked Prompt Modeling), as an additional stage in the pre-training scheme outlined in Fig. 3. The overall training process comprises three stages: (1) an LtR (**L**earn **t**o **R**ead) [22, 23] stage, (2) an LMPM stage and (3) a fine-tuning stage over the downstream task.

Learn to Read (LtR) The objective of this stage is to equip the model with the ability to comprehend text effectively. The crucial role of this step has been demonstrated in OCR-Free literature [22,23],where its significance in improving the model's ability to process textual information within documents has been shown. Hence, for this basic stage, we align with previous work and adopt the pre-training task of predicting the words in the document in their order of appearance, supervised by external annotations or OCR transcription of the text. The corresponding loss function reads:

$$\mathcal{L}_{LtR} = \mathcal{L}_{CE}\left(\mathcal{M}(X), Y_{OCR}\right) \tag{6}$$

where $\mathcal{L}_{CE}$ denotes the cross entropy loss, and Y_{OCR} is the top-to-bottom and left-to-right raster-scan order of the text in the input document X.

Fig. 4: Attention maps of the last ViLMA layer. Textual regions relevant to the question tokens are highly activated when performing LMPM pre-training (top) compared to not performing this training stage (bottom). It can be seen that the model focuses its attention not only on the specific input word but also on related words, e.g when performing cross attention with the word "diameter" it focuses on the words "under-ream" and "180 degrees".



Localized Masked Prompt Modeling (LMPM) The primary objective of this stage is to guide the model's attention towards the pertinent sections of the document. To address this, we leverage T5's [37] denoising objective: for a given sequence, a portion of the tokens are randomly masked and replaced by sentinel tokens, where spans of consecutive tokens are assigned to a single sentinel token. The task is to predict the omitted spans of tokens, separated by the same sentinel tokens utilized in the input sequence.

While adopting this general approach, instead of processing the entire document text, we opt to randomly sample a local span of tokens as illustrated on the right side of Fig. 3. We then apply the Masked Language Modeling (MLM) task to this span while keeping the visual text visible. This masked span $\mathbf{s}$ is subsequently used as a prompt, as we apply the cross entropy loss to predict the masked tokens:

$$\mathcal{L}_{LMPM} = \mathcal{L}_{CE}\left(\mathcal{M}(\mathbf{s}, X), Y_{LMPM}\right) \tag{7}$$

where Y_{LMPM} is the set of masked tokens, and $\mathbf{s}$ is the masked sampled sequence. Given that the LtR task remains constant and is not dependent on the prompt, this stage is trained with the original Swin patch merging layers. The integration of ViLMA layers into the architecture occurs only in the LMPM stage where those layers are randomly initialized.

While feeding the prompt embedding directly to the language model is the common practice, and has been shown to be effective [1,7,9,24,25], incorporating the prompt as input for both the vision encoder and the language model presents a challenge: the language model might compensate for any missing visual information from the vision encoder, effectively performing the Masked Language Modeling (MLM) task. To address this concern and encourage the vision encoder to develop its own focusing capabilities independently of the language model, both components need to be trained accordingly. To foster this independence, inspired by the Dropout technique [39], we adopt a strategy where the prompt is concatenated to the language model's input with a certain probability $\rho \in [0, 1]$ and omitted otherwise, such that,

$$\hat{\mathcal{Y}} = \begin{cases} \mathcal{M}_L\left(\mathbf{p}, \hat{Z}\right) & \text{if } \epsilon < \rho \\ \mathcal{M}_L\left(\hat{Z}\right) & otherwise \end{cases} \tag{8}$$

where $\epsilon \in [0, 1]$ is sampled uniformly at random at every training step. The benefits of applying LMPM are illustrated in Fig. 4 where the attention maps of the last ViLMA layer are depicted relative to words from the input query. In the top panel, where LMPM is utilized, the model exhibits a focused attention on relevant words, while in the bottom panel, where the LMPM stage is omitted, the model's attention appears scattered. For instance, when examining attention relative to the word "diameter" the model trained with LMPM focuses on related terms such as "under-ream" and "180 degrees" showcasing its improved ability to discern contextually relevant information. Further quantitative analysis of LMPM and Eq. (8) are demonstrated in Sec. 4.3.

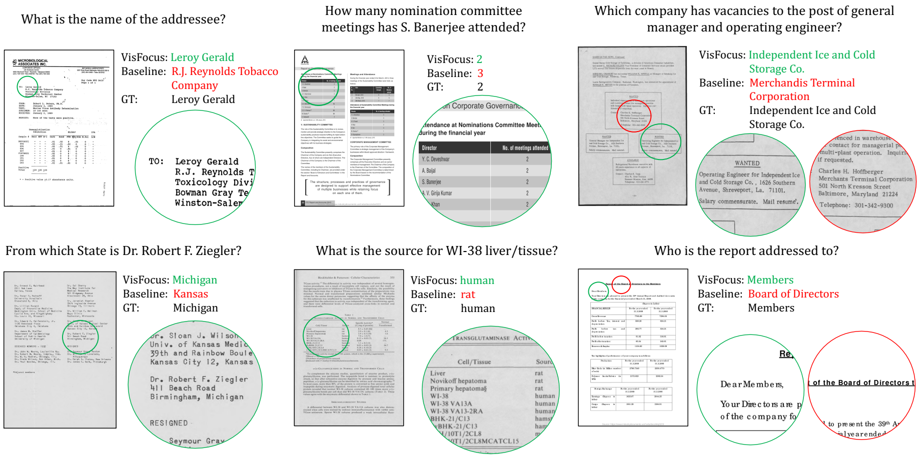
4 Experiments

We first present our experimental setup in Sec. 4.1, followed by results comparing to previous approaches on various Document VQA benchmarks in Sec. 4.2. We then present an ablation study showing the contribution individual components of our method and their synergy in Sec. 4.3. Finally we show a performance analysis VisFocus on varying document densities, showcasing its even increasing contribution for dense documents. Implementation details and further information on all datasets can be found in the appendix.

4.1 Experimental Setup

As stated in Sec. 3, VisFocus is composed of three main components: a vision encoder trained to extract visual features from high-resolution documents images, a projection module, and an Language Model (LM) that receives both the prompt and the projected visual features, yielding the final output.

Fig. 5: Qualitative comparison. We present examples from the DocVQA validation set, demonstrating our model's ability to accurately answer questions on denser documents compared to the baseline.


What is the name of the addressee?
VisFocus: Leroy Gerald
Baseline: R.J. Reynolds Tobacco Company
GT: Leroy Gerald
TO: Leroy Gerald
R.J. Reynolds T
Toxicology Divi
Bowman Gray Te
Winston-Saler
How many nomination committee meetings has S. Banerjee attended?
VisFocus: 2
Baseline: 3
GT: 2
Which company has vacancies to the post of general manager and operating engineer?
VisFocus: Independent Ice and Cold Storage Co.
Baseline: Merchandis Terminal Corporation
GT: Independent Ice and Cold Storage Co.
From which State is Dr. Robert F. Ziegler?
VisFocus: Michigan
Baseline: Kansas
GT: Michigan
Dr. Robert F. Ziegler
411 Beach Road
Birmingham, Michigan
RESIGNED
What is the source for WI-38 liver/tissue?
VisFocus: human
Baseline: rat
GT: human
Cell/Tissue
Source
Who is the report addressed to?
VisFocus: Members
Baseline: Board of Directors
GT: Members
Dear Members,

Implementation Details VisFocus utilizes SwinV2 [27] and T5 models [36] as the vision encoder and language model, pre-trained on ImageNet-1K [11] and "Colossal Clean Crawled Corpus" (C4) [37], respectively. In all our experiments, we utilize the SwinV2-Small variant. To align the output visual features with the LM, we employ a small multi-layer perceptron. This module, trained from scratch, projects the vision encoder's output into a shared latent space with the LM's input, before feeding it into the LM as input embeddings. We introduce two variants of our model: VisFocus-S and VisFocus-B, which incorporate the T5-Small and T5-Base variants, respectively. We compare VisFocus against corresponding baselines without the ViLMA layers and LMPM pre-training stage. We refer to those as "Baseline-{S,B}". In all our experiments, for both training and fine-tuning, we train on 8 A100 GPUs with bfloat-precision. We use AdamW [29] optimizer with cosine annealing [28] learning rate scheduler and warm-up. We train on high resolution input images of 1536×768. The complete implementation details and training recipes can be found in the Appendix.

Datasets and Metrics At the pre-training stages, we train our models on the IDL dataset [8] of document pages with OCR annotations. We evaluate our method against previous approaches over five different VQA benchmarks containing various domains: documents, infographics, charts and book covers. DocVQA [32], a subset of IDL, consists of $14k$ document images and $40k$ questions. InfographicsVQA (InfoVQA) [31] consists of $5k$ infographic images crawled from the web and $30k$ questions. ChartQA [30] contains chart images with questions requiring both visual and logical reasoning. It consists of $9.6K$ human-written questions and $23.1K$ generated questions based on summaries. OCR-VQA [33] is a large-scale dataset of $200k$ book cover images and $1M$ questions.

AI2 Diagrams (AI2D) [21] consists of $5K$ grade-school science diagrams, with corresponding multiple-choice questions. For DocVQA and InfoVQA we report Average Normalized Levenshtein Similarity (ANLS) metric [32], for ChartQA we follow [30] and report average Relaxed Accuracy (RA), and on OCR-VQA and AI2D we report Exact Match (EM). All of which are defined in the appendix for brevity. More information about datasets and metrics can found in Appendix.

Baselines We compare VisFocus to state-of-the-art OCR-free approaches for the small and base model size category. In the small category we compare to Dessurt [10] and Donut [22] which are pre-trained using the LtR task (Eq. (6)) on a corpus of real and synthetic documents. In the base category we compare to Pix2Struct-B [23] which uses screen-shot parsing as a pre-training task, and ScreenAI [4] which uses screen user interfaces for pre-training. For completeness, we present the results of other notable LVLMs.

4.2 Comparison to Previous Approaches

Table 1: Comparison with previous OCR-Free methods on VQA benchmarks. VisFocus outperforms previous methods of comparable scale, even when trained on substantially less pre-training data. We report ANLS on DocVQA and InfoVQA, Relaxed Accuracy (RA) on ChartQA and Exact Match (EM) on OCR-VQA and AI2D. In fully-trained methods, we only state total number of parameters.

	Method	#params (Trainable / Total)	DocVQA ANLS	InfoVQA ANLS	ChartQA RA	OCR-VQA EM	AI2D EM
Large	UReader [51]	86M / 7B	65.4	42.2	59.3	-	-
	Pix2Struct-L [23]	1.3B	76.6	40.0	58.6	**71.3**	42.1
	mPlugDocOwl [52]	7B	62.2	-	57.4	-	-
	PaLi-3 [9]	5B	**87.6**	**57.8**	**76.7**	70.0	**75.2**
Small	Dessurt [10]	127M	63.2	-	-	-	-
	Donut [22]	176M	67.5	11.6	41.8	66.0	-
	Baseline-S	110M	67.0	24.7	49.3	66.6	**42.7**
	VisFocus-S	132M / 171M	**68.6** (+1.6)	**28.5** (+3.8)	**53.0** (+3.7)	**67.3** (+0.7)	42.6 (-0.1)
Base	ScreenAI-B [4]	670M	50.7	19.6	54.0	54.8	-
	Pix2Struct-B [23]	282M	72.1	**38.2**	56.0	69.4	40.9
	Baseline-B	273M	71.7	26.8	52.5	66.9	45.6
	VisFocus-B	295M / 408M	**72.9** (+1.2)	31.9 (+5.1)	**57.1** (+4.6)	**70.0** (+3.1)	**47.8** (+2.2)

We compare the performance of our model to previous methods over five different benchmarks in Tab. 1 as specified in Sec. 4.1. The listed methods are categorized into three groups by model size. Those of base and small sizes are compared against our VisFocus variants of the same size category. Large models with billions of parameters are included for completeness, as those require substantially more data and computational resources. VisFocus improves over the baseline across all datasets in the base category and most datasets in the small category. It can be seen that VisFocus-S and VisFocus-B yield a performance gap of +**1.6**, +**1.2** points on DocVQA, +**3.8**, +**5.1** on InfoVQA, +**3.7**, +**4.6** on ChartQA, +**0.7**, +**3.1** on OCR-VQA and **-0.1**, +**2.2** on AI2D over their corresponding baselines. Notice that the additional parameter count attributed to the introduction of ViLMA layers are an order of magnitude smaller than the model size. Our approach achieves state-of-the-art performance, surpassing prior methods on four out of five benchmarks for the small category and on all benchmarks for the base category. While the performance of VisFocus over InfoVQA is significantly better than other methods oriented at equipping the model with reading capabilities (e.g., Donut and Dessurt), it is still lower than Pix2Struct. Considering that reasoning about infographics not only requires reading textual information but also processing other visuals, the gap is likely attributed to our focus on reading the most relevant parts of the document compared to Pix2Struct, which trains with more diverse pre-training tasks and over a larger diverse dataset, not publicly available.

4.3 Ablation Study and Empirical Analysis

Table 2: Breaking down the contributions of VisFocus' main components. ANLS for DocVQA and RA for ChartQA are reported.

Method	Prompt Interaction		LMPM Stage		DocVQA	ChartQA
	Concat	ViLMA	Concat	Alternate	ANLS	RA
Baseline-B	✓				70.9	52.5
+ViLMA	✓	✓			71.3	54.7
+LMPM	✓	✓	✓		71.8	55.7
+concat (Eq. (8)) (VisFocus-B)	✓	✓	✓	✓	**72.2**	**57.1**

We conduct an ablation study breaking down the impact of each component of VisFocus individually and pilling those up gradually to showcase the synergy when using both architectural enhancements and the pre-training scheme together. We evaluate VisFocus-B and report ANLS on the formal validation set of DocVQA for simplicity and the formal test set of ChartQA in Tab. 2. Each row in the table represents each of our contributions added independently, starting from our baseline, with all of the examined components disabled.

ViLMA We first quantify the contribution of architectural enhancements alone. As can be seen in the second line of Tab. 2, the transition from Swin's patch-merging layers to ViLMA layers add +**0.4** and +**2.2** points on DocVQA and ChartQA respectively.

LMPM The contribution of the ViLMA layers fulfills its potential when complemented by an appropriate pre-training task to encourage the encoded visual features to focus on relevant text patches. This is reflected by the additional +**0.5** and +**1.0** points on DocVQA and ChartQA respectively, as specified in the third row of Tab. 2. To ensure that the vision encoder attends to the prompt (provided via the ViLMA layers) and does not ignore it, we employ Eq. (8) to randomly skip the concatenation of the prompt to the LM's input. The final row in Tab. 2 quantifies the benefits of this technique by +**0.4** and +**1.4** point gains on DocVQA and ChartQA, respectively.

Table 3: Prompt Insertion Methods. Inserting the prompt via ViLMA layers improves results compared to previous approaches with only LtR pre-training applied (without LMPM). "Render"=question is rendered on the document image.

Injection Strategy	DocVQA ANLS	ChartQA RA
Baseline (LM-only)	70.9	52.5
Render (Pix2Struct)	70.6	52.2
VisFocus (ViLMA)	**71.3**	**54.7**

Table 4: ViLMA Layers Locations. Using ViLMA layers instead of all of the patch merging layers improves results compared to replacing part of the layers.

	Integration Stages	DocVQA ANLS	ChartQA RA
Baseline	`none`	70.9	52.5
VF-Early	`[1,2]`	71.0	54.1
VF-Mid	`[2,3]`	71.3	54.4
VF-Late	`[3,4]`	71.6	55.3
VF-All	`[1,2,3,4]`	**72.2**	**57.1**

Prompt Insertion Methods In this section we examine alternative ways to insert the prompt to the model, comparing our ViLMA layers to previous approaches. Tab. 3 compares between (1) the baseline approach of inserting the prompt to the language model alone (as done in e.g. Donut and Dessurt), (2) the approach suggested in Pix2Struct of rendering the prompt on top of the input image, and (3) our newly introduced ViLMA layers. ViLMA layers insert the prompt directly to the vision encoder patch-merging layers in addition to the LM input. This approach yields an improvement over the baseline, achieving gains +**0.4** and +**2.2** points on DocVQA and ChartQA, respectively. The rendering approach on contrast, lowers the results compared to the baseline approach. Note that for a fair comparison we pre-trained all of the models in the same way with LtR only (Eq. (6)) and without LMPM. Hence, we hypothesize that rendering the prompt on top of the image was more effective when applied under the pre-training tasks and data suggested in Pix2Struct, but underperforms when only fine-tuned using this approach.

To further quantify the contribution of substituting patch-merging layers with ViLMA layers within each block, we perform an ablation study as presented in Tab. 4. Each row of the table corresponds to a fine-tuning experiment where a subset of patch-merging layers are replaced with ViLMA layers. Our findings indicate that integrating ViLMA layers into deeper blocks yields more significant improvements. Specifically, employing ViLMA layers in all blocks results in the highest performance enhancement of +**1.3** and +**4.6** points, compared to a relatively lower improvement of +**0.7** and +**2.8** points on DocVQA and ChartQA respectively, when replacing only the last two layers.

Qualitative Analysis Fig. 5 provides qualitative examples comparing Baseline-B and VisFocus-B. In the Baseline-B, the visual tokens represent the entire document content rather than just what is relevant to the specific prompt. This sometimes leads to incorrect predictions, as the baseline model may extract an answer from irrelevant text patches. This effect is visualized in Fig. 4, where the cross-attention maps inside the ViLMA layers are plotted as heatmaps on top of the input image, showing the interactions between different tokens in the prompt and the visual patches. These visualizations offer clear insights into how

VisFocus learns to subjectively encode the document in relation to the given input query.

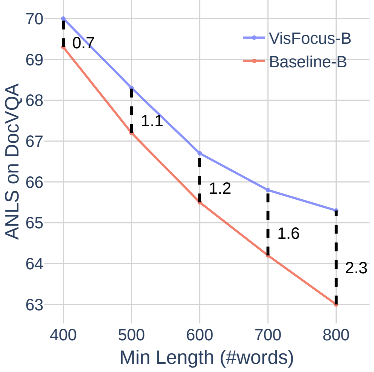


Fig. 6: Performance vs Number of Words. The graph shows the ANLS on a subset of the validation set of DocVQA containing at least Min Length words. The marginal gains achieved by VisFocus increase with the minimum number of words per document. This illustrates the significance of focusing the visual features on specific textual patches for dense documents.

Document Density Analysis In the following experiment we showcase the benefits of VisFocus's ability to focus on the most relevant text patches among all, possibly many, irrelevant text patches in dense documents. To this end, we measure the performance on subsets of documents with increasing densities (word counts). This is done by grouping the validation set of DocVQA to overlapping groups according to the minimum number of words (400, 500, 600, 700 and 800). For example, the first group consists of all documents containing at least 500 words, while the last group consists of all documents with at least 800 words. We compare the performance of VisFocus versus the baseline across these groups in Fig. 6. The consistently increasing performance gap on denser documents, ranging from +0.7 to +2.3 for all documents containing at least 400 and 800 words, respectively. This is aligned with our conjecture that in denser documents, focusing on the most relevant text patches to the specific user prompt is even more significant, given the larger amount of redundant information in the document.

5 Conclusions

In this work, we propose a novel way to make the visual encoding in OCR-Free VDU models aware of the user query. Consequently, the model learns to focus on reading the most relevant text in the document. The proposed method, VisFocus, couples the patch merging layers of a Swin transformer encoder with the user query inside newly introduced Vision-Language Merging Attention (ViLMa) layers. These are trained to focus mostly on encoding text relevant to the user query via a designated Localized Masked Prompt Modeling (LMPM) task. Those complementary components work in synergy to achieve state-of-the-art performance over a variety of document VQA tasks.

The purpose of this work is to equip the model with prompt-guided reading capabilities, and thus it is encouraged to focus on relevant text. A valid future research direction is the design of additional prompt-aware pre-train tasks, that guide the visual encoder to focus on content relevant to the user query beyond text. Specifically, this has the potential to improve performance on documents containing infographics, charts, and figures as well as on other domains.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022) 4, 9
2. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 993–1003 (2021) 2, 4
3. Appalaraju, S., Tang, P., Dong, Q., Sankaran, N., Zhou, Y., Manmatha, R.: Docformerv2: Local features for document understanding. In: AAAI Conference on Artificial Intelligence (2024) 2, 4
4. Baechler, G., Sunkara, S., Wang, M., Zubach, F., Mansoor, H., Etter, V., Cărbune, V., Lin, J., Chen, J., Sharma, A.: Screenai: A vision-language model for ui and infographics understanding. arXiv preprint arXiv:2402.04615 (2024) 4, 11
5. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 5
6. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023) 2
7. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023) 4, 9
8. Biten, A.F., Tito, R., Gomez, L., Valveny, E., Karatzas, D.: Ocr-idl: Ocr annotations for industry document library dataset. In: European Conference on Computer Vision. pp. 241–252. Springer (2022) 10
9. Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., Mustafa, B., Goodman, S., Alabdulmohsin, I., Padlewski, P., Salz, D., Xiong, X., Vlasic, D., Pavetic, F., Rong, K., Yu, T., Keysers, D., Zhai, X., Soricut, R.: Pali-3 vision language models: Smaller, faster, stronger (2023) 4, 9, 11
10. Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., Morariu, V.: End-to-end document recognition and understanding with dessurt. In: European Conference on Computer Vision. pp. 280–296. Springer (2022) 2, 3, 4, 11
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 10
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020), https://arxiv.org/abs/2010.11929 4
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020) 4

14. Fujinuma, Y., Varia, S., Sankaran, N., Appalaraju, S., Min, B., Vyas, Y.: A multimodal multilingual benchmark for document image classification. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 14361–14376. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.958, https://aclanthology.org/2023.findings-emnlp.958 1
15. Ganz, R., Kittenplon, Y., Aberdam, A., Avraham, E.B., Nuriel, O., Mazor, S., Litman, R.: Question aware vision transformer for multimodal reasoning. arXiv preprint arXiv:2402.05472 (2024) 4
16. Ganz, R., Nuriel, O., Aberdam, A., Kittenplon, Y., Mazor, S., Litman, R.: Towards models that can see and read. arXiv preprint arXiv:2301.07389 (2023) 6
17. Guillaume Jaume, Hazim Kemal Ekenel, J.P.T.: Funsd: A dataset for form understanding in noisy scanned documents. In: Accepted to ICDAR-OST (2019) 1
18. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval 1
19. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091 (2022) 2, 4
20. Hwang, W., Lee, H., Yim, J., Kim, G., Seo, M.: Cost-effective end-to-end information extraction for semi-structured document images. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3375–3383. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.271, https://aclanthology.org/2021.emnlp-main.271 2
21. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. ArXiv **abs/1603.07396** (2016), https://api.semanticscholar.org/CorpusID:2682274 11
22. Kim, G., Hong, T., Yim, M., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Donut: Document understanding transformer without ocr. arXiv preprint arXiv:2111.15664 **7**, 15 (2021) 2, 3, 4, 7, 11
23. Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., Eisenschlos, J.M., Khandelwal, U., Shaw, P., Chang, M.W., Toutanova, K.: Pix2struct: Screenshot parsing as pretraining for visual language understanding. In: International Conference on Machine Learning. pp. 18893–18912. PMLR (2023) 2, 4, 7, 11
24. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models (2023) 4, 5, 9
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023) 9
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023) 4, 5
27. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022) 6, 10
28. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with restarts. CoRR **abs/1608.03983** (2016), http://arxiv.org/abs/1608.03983 10
29. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. CoRR **abs/1711.05101** (2017), http://arxiv.org/abs/1711.05101 10
30. Masry, A., Long, D., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 2263–2279.

Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.177, https://aclanthology.org/2022.findings-acl.177 1, 10, 11

31. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022) 1, 10
32. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021) 1, 10, 11
33. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: ICDAR (2019) 10
34. Powalski, R., Łukasz Borchmann, Jurkiewicz, D., Dwojak, T., Pietruszka, M., Pałka, G.: Going full-tilt boogie on document understanding with text-image-layout transformer. ArXiv **abs/2102.09550** (2021), https://api.semanticscholar.org/CorpusID:231951453 4
35. Powalski, R., Łukasz Borchmann, Jurkiewicz, D., Dwojak, T., Pietruszka, M., Pałka, G.: Going full-tilt boogie on document understanding with text-image-layout transformer (2021) 2, 4
36. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv e-prints (2019) 10
37. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020) 8, 10
38. Seunghyun, P., Seung, S., Bado, L., Junyeop, L., Jaeheung, S., Minjoon, S., Hwalsuk, L.: Cord: A consolidated receipt dataset for post-ocr parsing (2019) 1
39. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014), http://jmlr.org/papers/v15/srivastava14a.html 9
40. Taghva, K., Beckley, R., Coombs, J.: The effects of ocr error on the extraction of private information. In: Document Analysis Systems VII: 7th International Workshop, DAS 2006, Nelson, New Zealand, February 13-15, 2006. Proceedings 7. pp. 348–357. Springer (2006) 2
41. Tanaka, R., Iki, T., Nishida, K., Saito, K., Suzuki, J.: Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions (2024) 2, 4
42. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C.Y., Bansal, M.: Unifying vision, text, and layout for universal document processing. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19254–19264 (2022), https://api.semanticscholar.org/CorpusID:254275326 2, 4
43. Tay, Y., Dehghani, M., Tran, V.Q., Garcia, X., Wei, J., Wang, X., Chung, H.W., Bahri, D., Schuster, T., Zheng, S., et al.: Ul2: Unifying language learning paradigms. In: The Eleventh International Conference on Learning Representations (2022) 5
44. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 5

45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 3, 4, 7
46. Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., Liu, X.: Docllm: A layout-aware generative language model for multimodal document understanding (2023) 2, 4
47. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175 (2023) 6
48. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2579–2591. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.201, https://aclanthology.org/2021.acl-long.201 2, 4
49. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200 (2020) 2, 4
50. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., et al.: mplug-docowl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499 (2023) 2, 4
51. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., et al.: Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023) 2, 11
52. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023) 2, 4, 11
53. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers (2021) 4
54. Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., Sun, T.: Llavar: Enhanced visual instruction tuning for text-rich image understanding (2023) 4, 5