Textual Grounding for Open-vocabulary Visual Information Extraction in Layout-diversified Documents

Mengjun Cheng[1,2], Chengquan Zhang[3], Chang Liu[4]✉, Yuke Li[1], Bohan Li[3], Kun Yao[3], Xiawu Zheng[5], Rongrong Ji[5], and Jie Chen[1,2]

[1] School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Department of Computer Vision Technology, Baidu Inc.
[4] Department of Automation, Tsinghua University, Beijing, China
[5] School of Information Science and Technology, Xiamen University, Fujian, China
liuchang2022@tsinghua.edu.cn; chenj@pcl.ac.cn

Abstract. Current methodologies have achieved notable success in the closed-set visual information extraction (VIE) task, while the exploration into open-vocabulary settings is comparatively underdeveloped, which is practical for individual users in terms of inferring information across documents of diverse types. Existing proposal solutions, including named entity recognition methods and large language model-based methods, fall short in processing the unlimited range of open-vocabulary keys and missing explicit layout modeling. This paper introduces a novel method for tackling the given challenge by transforming the process of categorizing text tokens into a task of locating regions based on given queries also called textual grounding. Particularly, we take this a step further by pairing open-vocabulary key language embedding with corresponding grounded text visual embedding. We design a document-tailored grounding framework by incorporating layout-aware context learning and document-tailored two-stage pre-training, which significantly improves the model's understanding of documents. Our method outperforms current proposal solutions on the SVRD benchmark for the open-vocabulary VIE task, offering lower costs and faster inference speed. Specifically, our method infers 20× faster than the QwenVL model and achieves an improvement of 24.3% in the F-score metric.

Keywords: Visual Informartion Extraction · Textual Grounding · Open-vocabulary

1 Introduction

The visual information extraction (VIE) task aims to extract corresponding information from visually-rich documents based on given query keys. Traditional researches, however, mainly focus on the evaluation of FUNSD [13], XFUNSD [42]

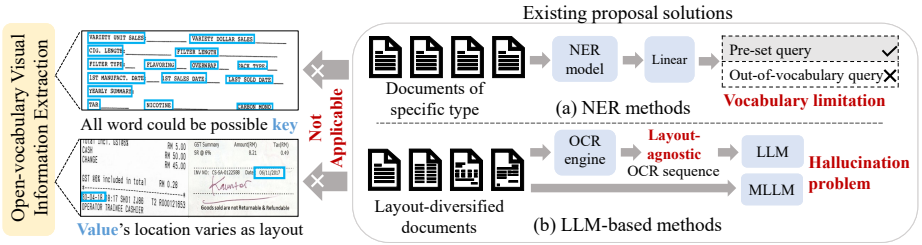

Fig. 1: The open-vocabulary VIE task requires models to have the capability for open-vocabulary query understanding and layout-aware context learning. However, existing approaches are not suitable to address this task. Specifically, (a) Named entity recognition (NER) methods are designed for pre-set queries and do not support out-of-vocabulary queries. (b) LLM-based methods use an OCR engine to extract text from layout-diversified documents, but their inputs are layout-agnostic.

etc. datasets, which model the information extraction task as named entity recognition (NER) task for pre-set queries like "question", "answer", and "header" and seen document types. They do not consider the requirements of individual users in the real world. Specifically, Individual users may provide arbitrary documents with unseen layouts and arbitrary query words that were not encountered during training. These situations expand the range of the VIE task from pre-set queries and seen documents to open-vocabulary queries and unseen document types with diversified layouts, which claims the capability of models for instruction understanding and layout-aware context learning.

However, existing proposal solutions including NER models and LLM-based models are not suitable to address this task. On the one hand, traditional approaches [2,11,40,49] for NER task model the extraction task as classification on pre-set queries. Their final linear layers are set fixed output dimensionality for pre-set queries, which means that they do not support reasonable prediction for new queries without training. Therefore, these NER methods struggle to encode out-of-vocabulary queries, resulting in limited capability for instruction understanding. On the other hand, while LLM-based approaches [14,20,27,35,56] show powerful understanding of open-vocabulary queries and achieve advances in the document question-answering task [3, 4, 36, 45, 53], they ignore learning contextual relationships between segments in different layouts. These approaches tend to classify segments according to their semantics rather than their contextual relationships, resulting in misjudgments among similar segments. In addition, LLM-based approaches sometimes generate hallucination words that do not appear in documents [21, 31], which is detrimental for the extraction task.

We address the open-vocabulary VIE task by redefining how we handle this task. Essentially, we transform the process of categorizing text tokens obtained from the OCR engine into a task of locating regions based on given queries, also called textual grounding. On the one hand, segments of documents are extracted explicitly in this scenario and fed to context learning according to their

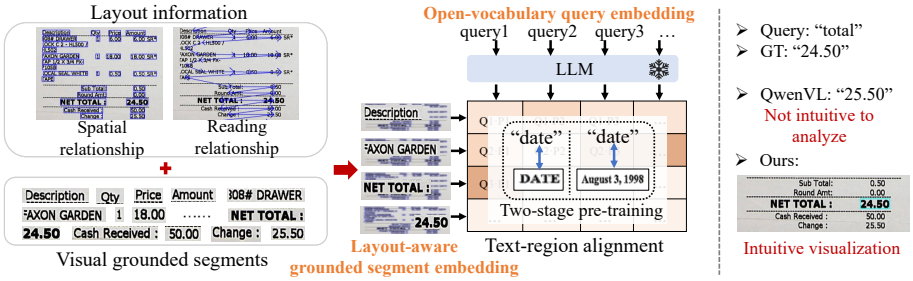

Fig. 2: We model the open-vocabulary VIE task by pairing the open-vocabulary query embedding and layout-aware grounded segment embedding for cross-modal text-region alignment. In this design, we extract information from documents by locating segments considering context information, making the predictions intuitive with visualization.

locations, which refers to the layout modeling. We use grounded text images to replace the non-differentiable OCR results, enabling end-to-end training and avoiding recognition errors from the OCR engine. On the other hand, open-vocabulary keys are treated as text queries and are encoded by a frozen LLM to obtain semantic embedding in language space. Finally, the open-vocabulary language semantic embedding and grounded text visual embedding are paired for text-region alignment, forming the foundation of our initial grounding architecture for the open-vocabulary VIE task.Under this architecture, multiple keys can be evaluated simultaneously, resulting in faster inference times compared to multiple rounds of question-answering. Additionally, this process of locating words avoids the possibility of generating hallucinations and the neglect of texts that are difficult to recognize and are sometimes ignored by traditional methods.

However, applying a pure textual grounding architecture, originally designed for natural images, directly to the field of document analysis proves to be unsuitable. Such an approach lacks the capability to grasp layout information and comprehend the semantics of visual words. Consequently, we further devise a layout-aware cross-modal architecture with grounded document pre-training objectives to enhance the model for document understanding. On the one hand, we design a coarse-to-fine document image encoding bench to extract layout-aware grounded segment embedding, which fuses image and language modalities at both the pixel and segment levels and integrates layout information. On the other hand, we design two-stage pre-training objectives based on text-region alignment for cross-modal semantic alignment and VIE modeling respectively.

In conclusion, this paper makes the following contributions:

- We model the open-vocabulary visual information extraction as the textual grounding task for the first time, aiming for open-vocabulary instruction understanding and layout-aware context learning.
- We design a document-tailored grounded architecture with layout-aware context learning and two-stage pre-training objectives for end-to-end training.

– Our method outperforms current proposal solutions on the SVRD benchmark for the open-vocabulary VIE task with faster inference speed. Specifically, our method infers 20× faster than the QwenVL model and achieves an improvement of about 24.3% in the F-score.

2 Related Work

2.1 Visual Information Extraction

Early approaches for the key information extraction (KIE) task heavily relied on extensive human-generated rules or patterns, which were often limited to specific types of documents and lacked generalizability [9, 25]. To move beyond these restrictive rules and patterns, subsequent studies have adopted learning-based methods using popular techniques like Convolutional Neural Networks (CNNs) [19, 54], Graph Convolutional Networks (GCNs) [47] and Transformers [24]. Compared to the KIE task, the VIE task focuses on retrieving information from visually-rich documents, particularly those containing diverse tables with varying styles and layouts. Therefore, the vision modality plays an important role in understanding the complicated layout of documents. Recently, pre-training models of Transformer architecture have become leading approaches in challenging NLP tasks and brought great performance improvement for the VIE task [2, 11, 16, 17, 22, 40, 41, 49]. LayoutLM series [11, 40, 41] are proposed to concurrently model image, text, and layout interactions in document images. GeolayoutLM [22] model geometric relationships explicitly in pre-training, achieving significant performance in entity linking from visually-rich documents. However, all of these models are developed for specific tasks and evaluated on closed-set benchmarks, which are not broadly applicable across different types of documents. A few works study a similar task, named the zero-shot key information extraction task, which urges the keys for evaluation are not exited in the train sets. KATA [5] addresses this task through a dual-stage architecture using text modality only, which first identifies the tagger of given keys in the document. IEMT [44] explores zero-shot keys in mixed-style tables by expanding the scale of pre-training for open-world learning. Those approaches may lead to failure when there are no trigger words in the document or corresponding keys in the vocabulary, which do not meet the requirement of open-vocabulary keys. Moreover, those approaches infer one image by multiple rounds for multiple keys like question-answering modeling, which reduces the search efficiency.

2.2 LLM-based Methods for Document Understanding

The rapid advancement of large language models (LLMs) has brought transformative changes to the field of artificial intelligence, which have demonstrated formidable capabilities in language understanding [26, 27, 33–35, 51, 52] through question-answering modeling. They could be also implemented for document understanding tasks by feeding document content, specifically, OCR results. Such

a strategy performs limited for visually-rich documents due to the lack of layout information and recognition errors from the OCR engine. Latin-prompt [38] attempts to feed layout information for LLMs by reconstructing the layout from OCR results, which achieves improvement to some extent for the document QA task. Recently, multimodal large language models (MLLMs) incorporate vision modality input for broader applications [1, 14, 20, 29, 46, 56]. Nevertheless, they often tend to generate hallucinations during text recognition and struggle to provide accurate predictions for the document understanding field [21, 31]. This limitation primarily stems from their inadequate layout modeling capabilities and handling of low-resolution inputs [21, 31]. Recently, some works take the document QA task, text recognition, or HTML parse into designs of pre-training objectives, achieving great performance on document understanding benchmarks [3, 4, 23, 36, 45, 53]. UReader [45] designs a shape-adaptive cropping module to leverage the frozen low-resolution vision encoder for processing high-resolution images. QwenVL [4] collects mounts of OCR data to improve the text-oriented tasks. Those models trained with document data achieve state-of-the-art performance among most document QA tasks but still struggle to achieve high accuracy for the information extraction task due to the absence of layout-aware context learning and the hallucination problem stemming from LLMs. These challenges highlight the need for continued research and development to enhance the performance of MLLMs in the domain of document understanding.

2.3 Textual Grounding

The textual grounding (TG) task aims to locate objects based on given queries. Existing methods are divided into two types: question-answering-based [7,8] and detection-based approaches [10, 15, 39, 50]. QA-based methods input the combination of text tokens and image tokens for transformer architecture and output a special token for following regression prediction [7,8]. Those methods are similar to LLM-based methods that need multiple rounds of inference to predict multiple keys of one document image, which is slow. Detection-based approaches stem from classical detection approaches and toward universal object detection for the real world. VILD [10] proposes open-vocabulary detection for the first time by utilizing the generalization of CLIP [30]. OV-DETR [50] addresses this task based on advanced DETR architecture. All those detection-based approaches remain the content of "class" and infer multiple open-vocabulary class names simultaneously. Our method follows the detection-based TG architecture and further proposes our document-oriented modification including the layout-aware context learning and document-tailored pre-training for document understanding.

3 Method

As illustrated in Fig. 3, our framework consists of three primary modules: the open-vocabulary query embedding module (discussed in Sec. 3.1), the layout-aware cross-modal embedding module (discussed in Sec. 3.2), and the text-region alignment module (discussed in Sec. 3.3).

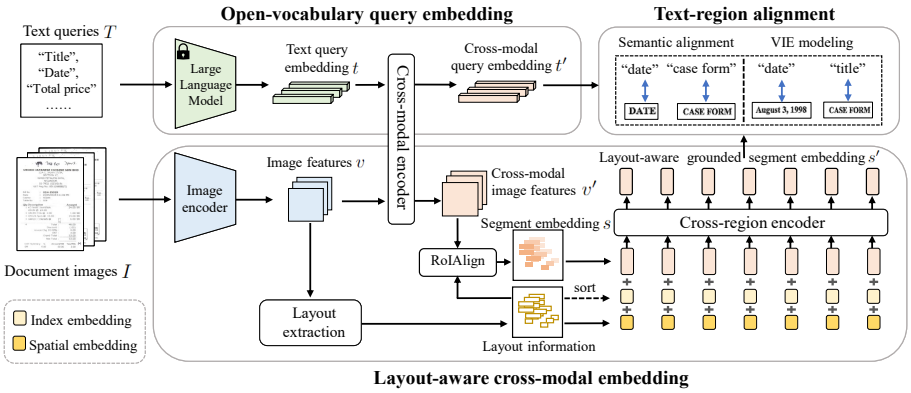

Fig. 3: Illustration of the proposed framework. We pair the cross-modal query embedding encoded from open-vocabulary keys with the layout-aware ground segment embedding into text-region pairs, and then conduct a two-stage pre-training for cross-modal semantic alignment and VIE modeling separately.

The first module encodes input text queries T to cross-modal query embedding t' for final region-text alignment through a frozen LLM and cross-modal encoder. The second module encodes input document images I into layout-aware grounded segment embedding s' for final alignment too. Specifically, preliminary image features v from the image encoder are calculated to generate layout information. The layout information combines encoded cross-modal image features v' to extract preliminary segment embedding s. The segment embedding is encoded by the cross-region encoder for layout-aware perception. Finally, we conduct two-stage pre-training between t' and s' for cross-modal semantic alignment and visual information extraction modeling respectively.

3.1 Open-vocabulary Query Embedding

Query Text Encoding. In the open-vocabulary VIE task, given text queries denoted as T, vary depending on the document type. Each text query has corresponding ground true text segments, forming key-value pairs. For each image, we select a fixed number of n query-segment pairs for training. Specifically, given text queries with their corresponding segments are sampled in a random sequence first, and then non-target segments with "background" captions are sampled as negative examples for completeness. Noted that, during the inference stage, given text queries of varying numbers are input in order without completeness.

Sampled text queries are tokenized into index sequences, represented as $e^T = [e_0^T, e_1^T, ..., e_{n-1}^T]$. Based on that, the text query embedding is then computed using a frozen large language model as

$$t = \Theta_{LLM}(\Theta_{\text{embedding}}(e^T)), \tag{1}$$

where $\Theta_{\text{embedding}}$ refers to the embedding layer of LLM, which projects index sequences into text tokens. The Θ_{LLM} encodes text tokens and then uses a global

pooling layer for each query to generate the text query embedding $t \in R^{n \times d}$, where d is the hidden dimensionality.

Cross-modal Embedding. To achieve document-aware embedding, we conduct cross-modal fusion between the text query embedding t and image features v. The flattened image features v are encoded from the input document image I via an image encoder at the pixel level. We concatenate those features denoted as $[v; t]$ and then input it into a cross-modal encoder. The cross-modal encoder comprises a stack of standard l_a transformer blocks with random initialization. The operational process of a transformer layer is formulated as

$$\begin{aligned} y_l &\leftarrow \mathrm{MHSA}(\mathrm{LN}([v_l; t_l])) + [v_l; t_l] \\ [v_{l+1}; t_{l+1}] &\leftarrow \mathrm{MLP}(\mathrm{LN}(y_l)) + y_l, \end{aligned} \tag{2}$$

where $\mathrm{MHSA}(\cdot)$ is the multi-head self-attention layer, $\mathrm{MLP}(\cdot)$ is the multi-layer perception layer, and $\mathrm{LN}(\cdot)$ denotes the layer normalization operator. The underlined signal indicates input features of the l-th layer. Following the cross-modal fusion, the output features of the final layer $[v_{l_a}; t_{l_a}]$ are divided into cross-modal image features v' and cross-modal query embedding t'. The latter one represents the semantic embedding of text queries specific to the document type.

3.2 Layout-aware Cross-modal Embedding

Overview. This module aims to get layout-aware grounded segment embedding, which refers to the proposal objects in the text-grounding task. We extract layout information and proposal segment embedding first and then feed layout information to segment embedding to get layout-aware segment embedding through the cross-region encoder. As a preliminary step, we calculate the input images I through an independent image encoder to get the preliminary image features $v \in R^{h \times w \times d}$ firstly for the following calculation. Here we use StrucTextv2 [49] as the image encoder for document images.

Layout Extraction. To effectively capture the spatial relationships between segments, we design a lightweight module for explicit layout information extraction. Specifically, we predict all segments within the input documents through a segmentation task, which is similar to [18]. Utilizing image features v encoded at a resolution of $(H/4, W/4)$, we employ a convolution layer with the 1×1 kernel to predict the map. The prediction map is upscaled to the original scale and then calculated to the probability map through a Sigmoid function. Subsequently, we extract the bounding box coordinates of predicted segments based on the binarized probability map, which is the layout information of documents.

Proposal Segment Embedding Extraction. In this step, we extract the preliminary proposal segment embedding s based on the cross-modal image features v' from the cross-modal encoder (introduced in Sec. 3.1). The segment

proposals are a combination of input ground truth segments and predicted segments. Specifically, we assign positive labels to the predicted segments through the Hungarian matching algorithm between them and input positive samples. The last predicted segments are assigned negative labels. Then we get sufficient negative samples for the following training. Subsequently, we extract the embedding of segment proposals on the cross-modal image features v' based on their locations through a RoIAlign operator. The segment embedding is obtained after an average pooling and denoted as $s = [s_0, s_1, ..., s_{m-1}] \in R^{m \times d}$, where m represents the total number of segment proposals.

Layout-aware Cross-region Segment Embedding. We augment the segment embedding with layout information using the cross-region encoder. The layout information is modeled by incorporating two types of position embedding: spatial embedding p^{sp} and index embedding p^{idx}. The spatial embedding p^{sp} represents the coordinate embedding of segments. Here, we encode the coordinates of the top-left and bottom-right corners of the segments' bounding boxes using trigonometric functions and sum the results. The index embedding p^{idx} indicates the sequence order of the segments. Proposal segments are arranged according to their coordinates in the reading order, specifically from top to bottom and from left to right. Flattened segment embedding s and two types of position embedding are summed as the input for the cross-modal encoder, formulated as

$$r_0 = [s_0, s_1, ..., s_{m-1}] + p^{idx} + p^{sp}, \tag{3}$$

where r_0 denotes the input embedding for the first layer of the cross-region encoder, which also consists of l_b transformer blocks with random initialization. The pipeline of layers in the cross-modal encoder is akin to that described in Eq. (2). The output embedding from the final layer of the cross-region encoder is referred to as the layout-aware grounded segment embedding $s' = [s'_0, s'_1, ..., s'_{m-1}]$. This embedding effectively integrates both the spatial and sequential aspects of the layout information, enhancing the overall performance of the model.

3.3 Text-region Alignment

Our approach implements segment understanding through the supervision of text-region alignment. As a preliminary step, we calculate the similarity between the normalized features of n cross-modal query embedding and m layout-aware grounded segment embedding by dot product, denoted as $C = \{c_{i,j}\}_{i=1,j=1}^{n,m}$. Here, $c_{i,j}$ represents the similarity score calculated between the i-th text embedding and the j-th vision embedding. These similarity scores are then utilized for cross-modal alignment across different objectives in our two-stage pre-training.

Cross-modal Semantic Alignment. The first stage of pre-training aims to enhance the model's ability to understand the semantics of words in the vision modality. This is achieved by aligning the semantics between words and their

corresponding visual words. Specifically, we sample segments from documents randomly, treating their OCR text as text queries for text-region alignment. So a text-region pair consists of OCR texts and their visual counterparts, such as the word "date" and its corresponding image crop as shown in Fig. 3. In this stage, we do not use predicted segments to augment the segment pool since all segments have their text. Hence segment proposals are n ground truth segments directly, resulting in a similarity matrix of size $n \times n$. We supervise this process using a standard contrastive loss. The text-to-image loss is formulated as

$$\mathcal{L}t2i = -\sum i \in \mathcal{B} \log \frac{\exp(c_{i,i}/\sigma)}{\sum_{j=1}^{N} \exp(c_{i,j}/\sigma)}, \tag{4}$$

where σ is a trainable temperature parameter initialized with 0.07 following [30]. $\mathcal{B}$ denotes the set of indexes of segment proposals. Same as the image-to-text loss. This training stage ensures that the encoded features of images with text are semantically linked to their corresponding text, paving the way for our OCR-free architecture.

Visual Information Extraction Modeling. The second stage supervises the VIE task by aligning text-region pairs of key queries and their corresponding values in the vision modality as shown in Fig. 3. We treat the given keys as text queries and get a similarity matrix of size $n \times m$ between n text queries and m segment proposals. Given that multiple segments can match the same text query, standard contrastive loss is not applicable. Drawing inspiration from [43], we employ a unified image-text-label contrastive loss to supervise the alignment of text-region pairs. Scores belonging to the same text query are grouped. The text-to-image loss is formulated as

$$\mathcal{L}_{t2i} = -\sum_{i \in \mathcal{B}} \frac{1}{|\mathcal{P}(i)|} \sum_{k \in \mathcal{P}(i)} \log \frac{\exp\left(c_{i,k}/\sigma\right)}{\sum_{j \in \mathcal{B}} \exp\left(c_{i,j}/\sigma\right)}, \tag{5}$$

where $k \in \mathcal{P}(i) = \{k \mid k \in \mathcal{B}, y_k = y_i\}$ identifies the set of segments that share the same text query as the i-th segment. Here, y_i signifies the label of the i-th segment, another form of text query. The image-to-text loss is similarly structured and is formulated as

$$\mathcal{L}_{i2t} = -\sum_{j \in \mathcal{B}} \frac{1}{|\mathcal{P}(j)|} \sum_{k \in \mathcal{P}(j)} \log \frac{\exp\left(c_{k,j}/\sigma\right)}{\sum_{i \in \mathcal{B}} \exp\left(c_{i,j}/\sigma\right)}. \tag{6}$$

This formulation of the image-text-label contrastive loss could be perceived as a multiple-label classification loss.

3.4 Discussion

We model the open-vocabulary visual information extraction task as a textual grounding task for the first time and construct a document-tailored grounding

architecture. Under this architecture, we achieve an end-to-end OCR-free training pattern by replacing traditional text tokens from the offline OCR engine with grounded segment images for online text-region alignment training. It shares the same goal as those document-related models using the OCR embedding from the OCR engine directly, which avoids recognition errors stemming from the OCR engine to some extent. We pursue fast inference speed, as processing pre-set keys can be repetitive when individual users need to use AI to handle large amounts of private documents. In addition, the primary focus of this paper is on demonstrating the effectiveness of the established pipeline based on the textual grounding solution. For some difficult scenarios such as poor illumination and font transformation, the layout extraction module needs to be more robust and intricately designed.

4 Experiments

4.1 Implementation Details

Backbone. We use the StructTextV2 [49] as the image encoder and ERNIE 3.0 [33] as the frozen language model. The rest of the model components are initially randomly initialized before the pre-training phase. The model is trained with a batch size of 16 and a learning rate of 5e-5. The dimensionality of the embeddings is set to 768. Both the cross-modal encoder and the cross-region encoder consist of 3 layers. We use dice loss and sigmoid focal loss to supervise the segmentation task for segment prediction.

Datasets. For the initial stage of pre-training, we collect data comprising private 350,000 Chinese documents and 300,000 English documents from the IIT-CDIP [32], which are large-scale scanned document image datasets. In the second stage, we incorporate a variety of open-source VIE datasets, including FUNSD [13], XFUND [42], CORD [28], SROIE [12], EPHOIE [37], SVRD [48], Publaynet [55], and M^6Doc [6]. A mini set from Publaynet and M^6Doc is selected for training purposes. We use data of SVRD task1 and task2 and the train set of task3, all of them have different document types with the evaluation set. The total dataset for VIE modeling consists of 11,456 document images, which is considerably smaller than the dataset used in the first pre-training stage or the dataset size typically used for large language models. Our evaluation of the method and comparison with other large models is conducted on the SVRD task3 test set, which includes ten document types such as tickets, bills, cards, certificates, mail, etc. Each document type is associated with its specific set of text queries, varying across different document types.

Metrics. We adopt the F-score metric following [48]. To compare with models that output language during the evaluation, we process the segments predicted by our method through an open-source API for OCR recognition, which converts

Table 1: Comparisons on the SVRD benchmark.

Model	Time(s)	bank	license	taxi	shop	vehicle	weight	ticket	bus	memo	paper	All
CLIP-VG$_{tuned}$ [39]	0.38	0.00	1.36	1.33	0.00	6.52	3.20	1.16	2.24	0.00	0.00	1.75
LayoutLMv2[1][40]	3.78	2.10	0.77	3.20	15.87	7.65	3.03	5.86	6.09	43.57	11.55	10.82
LLaMA2-7B [35]	1.90	0.00	0.77	0.80	0.00	3.66	1.89	0.73	1.17	0.37	3.03	1.23
LLaMA2-13B [35]	12.30	0.00	2.54	1.33	0.00	6.52	3.20	1.16	2.24	2.42	0.00	2.09
Qwen [3]	8.66	22.10	14.31	10.8	9.71	6.32	8.33	0.73	7.26	13.78	1.82	9.24
BLIP-2 [14]	3.36	0.00	0.77	0.80	0.00	3.32	1.89	0.73	1.17	0.00	0.00	0.96
UReader [45]	4.32	18.94	2.32	9.20	18.72	10.64	18.93	13.18	7.73	28.30	17.62	14.07
QwenVL [4]	5.33	26.31	4.64	8.00	27.48	13.97	16.28	16.11	20.63	21.97	17.62	17.63
QwenVL$_{tuned}$ [4]	18.60	43.15	22.97	7.20	33.88	25.95	19.31	30.76	32.35	32.40	**26.13**	27.68
TG-VIE (Ours)	**0.25**	**53.84**	**54.70**	**33.25**	**35.82**	**52.61**	**23.23**	**69.29**	**44.72**	**49.68**	7.74	**41.98**

[1] Inference by question-answering modeling.

these segments into language outputs. In the ablation studies, we employ the H-mean metric to assess models' performance, thereby minimizing the effect of recognition errors introduced by the OCR recognition model potentially.

4.2 Performance

We compare the performance of different approaches including LLMs, multi-modal LLMs, textual grounding method, and document pre-training method as shown in Tab. 1. The pure textual grounding approach, CLIP-VG, performs limited in this document field's task, even when fine-tuned with document data. Traditional NER method LayoutLMv2 can be inferred by QA modeling here, but it still performs not well in this task. LLM-based models share the same situations, even though they are trained on large-scale data and can comprehend open-vocabulary text queries. Even models trained with document data, such as Qwen, QwenVL, and UReader, perform far worse than our method, with a gap of more than 24%. This underperformance may be attributed to the small input resolution of these models and the absence of a layout-aware module in their design. The QwenVL model, fine-tuned with our training set, shows some improvement but remains insufficient. Our TG-VIE model performs well among all compared methods and infers quickly. It has a limited performance on the paper subset due to its different granularity of answer text from the training set, which requires a more advanced OCR recognition module to process. To facilitate the comparison, we provided each LLM-based model with a prefixed task description and a query template, such as "What's the { } ?". For pure LLMs, we additionally supply official OCR results to represent the document content. Further details are available in the supplementary materials.

Table 2: Impact of layout information.

Model	Setting	Precision	Recall	H-mean
Ours		51.71	50.01	50.58
	w/o index embed.	50.87	47.89	48.95
	w/o spatial embed.	48.66	45.32	46.58
	w/o layout info.	44.34	47.60	45.58

Table 3: Impact of segmentation losses.

$\mathcal{L}_{dice}$	$\mathcal{L}_{mask}$	Precision	Recall	H-mean
✓		37.49	39.50	38.33
	✓	28.64	30.71	29.40
✓	✓	38.50	41.14	39.52

Table 4: Impact of two-stage pre-training.

Setting	Precision	Recall	H-mean
VIE modeling only	38.60	41.14	39.52
+ cross-modal alignment	51.71	50.01	50.58

Table 5: Impact of sampled examples.

Model	Sample number				w/o labeled
	4	8	16	32	proposals
Ours	45.94	45.49	50.58	48.79	44.00

4.3 Ablations

Inference Speed. Compared to methods including LLM-based methods and the NER method that treat the VIE task as a question-answering task, our framework infers multiple queries simultaneously for each image. This design results in a significant speed advantage, for example, being 20× faster than QwenVL and 34× faster than Qwen, as indicated in Tab. 1. Note that, we query QA modeling methods with only one key at a time and repeat this process for each of the given keys. If we query them with multiple keys in one question, the performance of these LLM-based methods would likely decrease due to processing long tokens input. Additionally, our timing considerations are based solely on the inference time with processed input. If we take the time required by the preliminary OCR engine into account, the total time cost of LLMs would be even greater.

Impact of Layout Information. We explored the significance of two types of position embedding fed into the cross-region encoder by setting one as the zero vector. As Tab. 2 demonstrates, our framework benefits from both spatial and index embeddings to varying extents. The spatial information of segments within documents proves to be more critical than another one. Given the complex layout of visually-rich documents, a standard reading order is insufficient for comprehensively learning the context of segments. Eliminating all layout information severely impacts performance, as the model then relies on semantic alignment between text queries and segments for cross-modal understanding only.

Impact of Segmentation Losses. The accuracy of layout extraction is pivotal for the final prediction, as the recall rate of segment proposals sets the upper bound for matching accuracy in subsequent steps. We present the impact of different losses on performance in Tab. 3. It is evident that the Dice loss plays a more critical role compared to the other loss.

Table 6: Impact of explicit segment extraction for layout modeling.

Segment extraction	Precision	Recall	H-mean
	17.01	17.29	16.99
✓	38.50	41.14	39.52

Table 7: Impact of fusion level.

Pixel level	Segment Level	Precision	Recall	H-mean
✓		36.18	38.61	37.12
	✓	33.16	35.42	34.06
✓	✓	38.50	41.14	39.52

Impact of Two-stage Pre-training. The cross-modal semantic alignment pre-training significantly enhances our model's ability to learn the semantics of visual words, thereby aiding in the contextual understanding across segments. This process narrows the gap between the image embeddings and the text embeddings from the large language model within the hidden feature space. Tab. 4 reveals that omitting this pre-training step leads to an approximate 11% decrease in model performance according to the H-mean metric, underscoring the importance of cross-modal semantic alignment.

Impact of Sampled Examples. The number of sampled examples affects performance as shown in Tab. 5. Insufficient sampling fails to provide enough pairs for effective training. The optimal number of samples is closely tied to the distribution of key-value pairs during the training phase. Labeling segment proposals generates additional negative samples, which helps differentiate between target and non-target segments, consequently enhancing our model's performance. Without including any negative samples, the training process becomes considerably lengthier and the final performance is adversely affected.

Impact of Explicit Segment Extraction. We evaluated the criticality of explicitly extracting segment proposals by comparing our framework against a modified version that omits this step. Specifically, in the layout extraction module of the alternative framework, the supervised regions of the predicted map were adjusted from encompassing all segment proposals to only include target segments. This modification necessitates that the prediction is derived from cross-modal image features directly, as it relies on the text query for context. To facilitate a fair comparison, given the architectural changes, we conducted experiments without the first-stage pre-training. The results, as presented in Tab. 6, indicate a substantial decline in performance, with a roughly 22% decrease in the H-mean metric. This significant drop highlights the challenges faced by the framework without the contextual semantic learning between segment proposals in addressing open-vocabulary VIE tasks, which is particularly crucial for effectively modeling visually-rich documents.

Impact of Fusion at Different Levels. We fuse text features and image features in the cross-modal encoder for pixel-level fusion and construct layout-aware context learning in the cross-region encoder for segment-level. As evidenced in

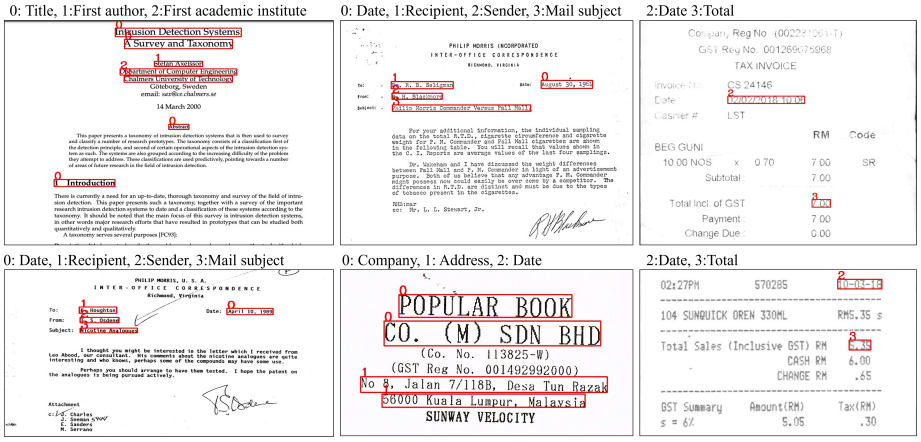

Fig. 4: Examples of various types of documents. The red boxes indicate matched segments with given keys.

Tab. 7, the removal of any encoder component has a detrimental effect on the model's performance. Specifically, omitting the cross-region encoder leads to an architecture that fails to perceive layout information, although it retains the positional information of vision patches. Additionally, an architecture that solely relies on the cross-modal encoder also encounters challenges. In this setup, the model struggles due to the absence of information outside the segment regions, and the vision embedding does not adequately concentrate on potentially relevant information in light of the given text queries. The results highlight the critical role of each encoder in comprehensive processing and document image understanding in our model.

Qualitative Comparisons. We show some examples predicted by our method on various types of documents in Fig. 4. More visualization is shown in the supplementary materials.

5 Conclusion

We have developed an innovative approach to address the open-vocabulary visual information extraction task across different documents with diversified layouts. This method reimagines the traditional process of categorizing text tokens from OCR outputs, transforming it into a task of locating regions based on the given text query, also known as textual grounding. We propose a coarse-to-fine cross-modal framework, complemented by layout-aware context learning and document-tailored two-stage pre-training, specifically tailored for document understanding. Experimental results show our framework exceeds existing proposal solutions including LLMs and multimodal LLMs in performance and efficiency.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0118201), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465), the Shenzhen Medical Research Funds in China (No. B2302037), and AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China.

References

1. Alayrac, J., 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.L., 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. In: Advances in Neural Information Processing Systems, NeurIPS 2022 (2022)
2. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: DocFormer: End-to-end transformer for document understanding. In: IEEE/CVF International Conference on Computer Vision, ICCV 2021. pp. 973–983. IEEE (2021)
3. 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. CoRR abs/2309.16609 (2023)
4. 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. CoRR abs/2308.12966 (2023)
5. Cao, R., Luo, P.: Extracting zero-shot structured information from form-like documents: Pretraining with keys and triggers. In: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2021. pp. 12612–12620. AAAI Press (2021)
6. Cheng, H., Zhang, P., Wu, S., Zhang, J., Zhu, Q., Xie, Z., Li, J., Ding, K., Jin, L.: M^6Doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 15138–15147. IEEE (2023)
7. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: TransVG: End-to-end visual grounding with transformers. In: IEEE/CVF International Conference on Computer Vision, ICCV 2021. pp. 1749–1759. IEEE (2021)
8. Deng, J., Yang, Z., Liu, D., Chen, T., Zhou, W., Zhang, Y., Li, H., Ouyang, W.: TransVG++: End-to-end visual grounding with language conditioned vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(11), 13636–13652 (2023)
9. Gentile, A.L., Zhang, Z., Ciravegna, F.: Web scale information extraction with LODIE. In: 2013 AAAI Fall Symposia Series. AAAI Press (2013)
10. Gu, X., Lin, T., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: The Tenth International Conference on Learning Representations, ICLR 2022. OpenReview.net (2022)

11. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: Pre-training for document AI with unified text and image masking. In: Magalhães, J., Bimbo, A.D., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., Toni, L. (eds.) Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091. ACM (2022)
12. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.V.: ICDAR2019 competition on scanned receipt OCR and information extraction. In: International Conference on Document Analysis and Recognition, ICDAR 2019. pp. 1516–1520. IEEE (2019)
13. Jaume, G., Ekenel, H.K., Thiran, J.: FUNSD: A dataset for form understanding in noisy scanned documents. In: 2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019. pp. 1–6. IEEE (2019)
14. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) International Conference on Machine Learning, ICML 2023. vol. 202, pp. 19730–19742. PMLR (2023)
15. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J., Chang, K., Gao, J.: Grounded language-image pre-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. pp. 10955–10965. IEEE (2022)
16. Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., Ding, E.: StrucTexT: Structured text understanding with multi-modal transformers. In: Shen, H.T., Zhuang, Y., Smith, J.R., Yang, Y., César, P., Metze, F., Prabhakaran, B. (eds.) Proceedings of the 29th ACM International Conference on Multimedia. pp. 1912–1920. ACM (2021)
17. Liao, H., RoyChowdhury, A., Li, W., Bansal, A., Zhang, Y., Tu, Z., Satzoda, R.K., Manmatha, R., Mahadevan, V.: Doctr: Document transformer for structured information extraction in documents. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023. pp. 19527–19537. IEEE (2023)
18. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(1), 919–931 (2023)
19. Lin, B.Y., Sheng, Y., Vo, N., Tata, S.: FreeDOM: A transferable neural architecture for structured information extraction on web documents. In: Gupta, R., Liu, Y., Tang, J., Prakash, B.A. (eds.) Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1092–1102. ACM (2020)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems, NeurIPS 2023 (2023)
21. Liu, Y., Li, Z., Li, H., Yu, W., Huang, M., Peng, D., Liu, M., Chen, M., Li, C., Jin, L., Bai, X.: On the hidden mystery of OCR in large multimodal models. CoRR abs/2305.07895 (2023)
22. Luo, C., Cheng, C., Zheng, Q., Yao, C.: GeoLayoutLM: Geometric pre-training for visual information extraction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 7092–7101. IEEE (2023)
23. Lv, T., Huang, Y., Chen, J., Cui, L., Ma, S., Chang, Y., Huang, S., Wang, W., Dong, L., Luo, W., Wu, S., Wang, G., Zhang, C., Wei, F.: Kosmos-2.5: A multimodal literate model. CoRR abs/2309.11419 (2023)

24. Majumder, B.P., Potti, N., Tata, S., Wendt, J.B., Zhao, Q., Najork, M.: Representation learning for information extraction from form-like documents. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. pp. 6495–6504. Association for Computational Linguistics (2020)
25. Medvet, E., Bartoli, A., Davanzo, G.: A probabilistic approach to printed document understanding. International Journal on Document Analysis and Recognition (IJDAR) 14(4), 335–347 (2011)
26. OpenAI: ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat (2023)
27. OpenAI: GPT-4 technical report. CoRR abs/2303.08774 (2023)
28. Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., Lee, H.: CORD: A consolidated receipt dataset for post-ocr parsing. In: Workshop on Document Intelligence at NeurIPS 2019 (2019)
29. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. In: The Eleventh International Conference on Learning Representations, ICLR 2024. OpenReview.net (2024)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
31. Shi, Y., Peng, D., Liao, W., Lin, Z., Chen, X., Liu, C., Zhang, Y., Jin, L.: Exploring OCR capabilities of GPT-4V(ision) : A quantitative and in-depth evaluation. CoRR abs/2310.16809 (2023)
32. Soboroff, I.: Complex document information processing (cdip) dataset. National Institute of Standards and Technology (2022)
33. Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., Liu, W., Wu, Z., Gong, W., Liang, J., Shang, Z., Sun, P., Liu, W., Ouyang, X., Yu, D., Tian, H., Wu, H., Wang, H.: ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. CoRR abs/2107.02137 (2021)
34. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and efficient foundation language models. CoRR abs/2302.13971 (2023)
35. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. CoRR abs/2307.09288 (2023)

36. 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. CoRR abs/2401.00908 (2024)
37. Wang, J., Liu, C., Jin, L., Tang, G., Zhang, J., Zhang, S., Wang, Q., Wu, Y., Cai, M.: Towards robust visual information extraction in real world: New dataset and novel solution. In: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2021. pp. 2738–2745. AAAI Press (2021)
38. Wang, W., Li, Y., Ou, Y., Zhang, Y.: Layout and task aware instruction prompt for zero-shot document image question answering. CoRR abs/2306.00526 (2023)
39. Xiao, L., Yang, X., Peng, F., Yan, M., Wang, Y., Xu, C.: CLIP-VG: Self-paced curriculum adapting of CLIP for visual grounding. IEEE Transactions on Multimedia 26, 4334–4347 (2024)
40. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florêncio, D.A.F., 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, ACL/IJCNLP 2021, (Volume 1: Long Papers). pp. 2579–2591. Association for Computational Linguistics (2021)
41. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of text and layout for document image understanding. In: Gupta, R., Liu, Y., Tang, J., Prakash, B.A. (eds.) Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery data mining. pp. 1192–1200. ACM (2020)
42. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florêncio, D.A.F., Zhang, C., Wei, F.: XFUND: A benchmark dataset for multilingual visually rich form understanding. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 3214–3224. Association for Computational Linguistics (2022)
43. Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J.: Unified contrastive learning in image-text-label space. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. pp. 19141–19151. IEEE (2022)
44. Yang, Q., Hu, Y., Cao, R., Li, H., Luo, P.: Zero-shot key information extraction from mixed-style tables: Pre-training on wikipedia. In: Bailey, J., Miettinen, P., Koh, Y.S., Tao, D., Wu, X. (eds.) IEEE International Conference on Data Mining, ICDM 2021. pp. 1451–1456. IEEE (2021)
45. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., Jin, Q., He, L., Lin, X., Huang, F.: Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 2841–2858. Association for Computational Linguistics (2023)
46. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., Huang, F.: mPLUG-Owl: Modularization empowers large language models with multimodality. CoRR abs/2304.14178 (2023)
47. Yu, W., Lu, N., Qi, X., Gong, P., Xiao, R.: PICK: Processing key information extraction from documents using improved graph learning-convolutional networks. In: 25th International Conference on Pattern Recognition, ICPR 2020. pp. 4363–4370. IEEE (2020)
48. Yu, W., Zhang, C., Cao, H., Hua, W., Li, B., Chen, H., Liu, M., Chen, M., Kuang, J., Cheng, M., Du, Y., Feng, S., Hu, X., Lyu, P., Yao, K., Yu, Y., Liu, Y., Che, W.,

Ding, E., Liu, C., Luo, J., Yan, S., Zhang, M., Karatzas, D., Sun, X., Wang, J., Bai, X.: ICDAR 2023 competition on structured text extraction from visually-rich document images. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) International Conference on Document Analysis and Recognition, ICDAR 2023. Lecture Notes in Computer Science, vol. 14188, pp. 536–552. Springer (2023)
49. Yu, Y., Li, Y., Zhang, C., Zhang, X., Guo, Z., Qin, X., Yao, K., Han, J., Ding, E., Wang, J.: StrucTexTv2: Masked visual-textual prediction for document image pre-training. In: The Eleventh International Conference on Learning Representations, ICLR 2023. OpenReview.net (2023)
50. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary DETR with conditional matching. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) European Conference on Computer Vision, ECCV 2022. vol. 13669, pp. 106–122. Springer (2022)
51. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W.L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Liu, Z., Zhang, P., Dong, Y., Tang, J.: GLM-130B: An open bilingual pre-trained model. In: The Eleventh International Conference on Learning Representations, ICLR 2023. OpenReview.net (2023)
52. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M.T., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: OPT: Open pre-trained transformer language models. CoRR abs/2205.01068 (2022)
53. Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., Sun, T.: LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. CoRR abs/2306.17107 (2023)
54. Zhao, X., Wu, Z., Wang, X.: CUTIE: Learning to understand documents with convolutional universal text information extractor. CoRR abs/1903.12363 (2019)
55. Zhong, X., Tang, J., Jimeno-Yepes, A.: PubLayNet: Largest dataset ever for document layout analysis. In: International Conference on Document Analysis and Recognition, ICDAR 2019. pp. 1015–1022. IEEE (2019)
56. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: The Eleventh International Conference on Learning Representations, ICLR 2024. OpenReview.net (2024)