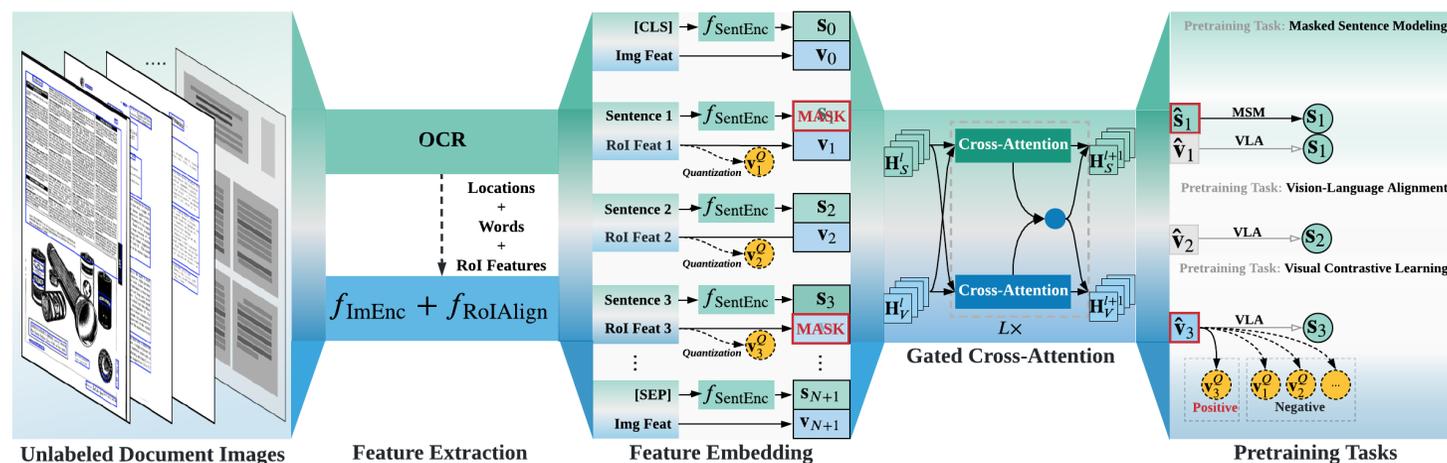


Unified Pretraining Framework for Document Understanding

**Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao,
Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, Tong Sun**

**Adobe Research,
Adobe Document Cloud**

Problem and Contribution



1. We introduce **UniDoc**, a powerful pretraining **framework** for document understanding. **UniDoc** is capable of learning contextual **textual and visual** information and **cross-modal correlations** within a **single framework**, which leads to better performance.
2. We present **Masked Sentence Modeling** for language modeling, **Visual Contrastive Learning** for vision modeling, and **Vision-Language Alignment** for pretraining.
3. Extensive experiments and analysis provide useful insights on the effectiveness of the pretraining tasks and show outstanding performance on various downstream tasks.

Motivation

1. Documents are composed of semantic regions

UniDoc: Unified Pretraining Framework for Document Understanding

Jiuxiang Gu¹, Jason Kuen¹, Vlad I. Morariu¹, Handong Zhao¹, Nikolaos Bampalios², Rajiv Jain¹, Ani Nenkova¹, Tong Sun¹
¹Adobe Research, ²Adobe Document Cloud

{jigu, kuen, morariu, hazhao, bampali, rajivjain, nenkova, tsun}@adobe.com

Abstract

Document intelligence automates the extraction of information from documents and supports many business applications. Recent self-supervised learning methods on large-scale unlabeled document datasets have opened up promising directions towards reducing annotation efforts by training models with self-supervised objectives. However, most of the existing document pretraining methods are still language-dominated. We present UniDoc, a new unified pretraining framework for document understanding. UniDoc is designed to support most document understanding tasks, extending the Transformer to take multimodal embeddings as input. Each input element is composed of words and visual features from a semantic region of the input document image. An important feature of UniDoc is that it learns a generic representation by making use of three self-supervised losses, encouraging the representation to model sentences, learn similarities, and align modalities. Extensive empirical analysis demonstrates that the pretraining procedure learns better joint representations and leads to improvements in downstream tasks.

1 Introduction

Document intelligence is a broad research area that includes techniques for information extraction and understanding. Unlike plain-text documents in natural language processing (NLP) [1, 2, 3], a physical document can be composed of multiple elements: tables, figures, charts, etc. In addition, a document usually includes rich visual information, and can be one of various types of documents (scientific paper, form, resume, etc.), with various combinations of multiple elements and layouts. Complex content and layout, noisy data, font and style variations make automatic document understanding very challenging. For example, to understand text-rich documents such as letters, a system needs to focus almost exclusively on text content, paying attention to a long sequential context, while processing semi-structured documents such as forms requires the system to analyze spatially distributed short words, paying particular attention to the spatial arrangement of the words. Following the success of BERT [4] on NLP tasks, there has been growing interest in developing pretraining methods for document understanding [5, 6, 7, 8]. Pretrained models have achieved state-of-the-art (SoTA) performance across diverse document understanding tasks [9].

Image training datasets help pretraining models to learn a good representation for downstream tasks; however, we observe three major problems with the current pretraining setup: (1) documents are composed of semantic regions. Most of the recent document pretraining works follow BERT and split documents into words. However, unlike the sequence-to-sequence learning in NLP, documents have hierarchical structure (words form sentences, sentences form a semantic region, and semantic region form a document). Also, the importance of words and sentences are highly context-dependent, i.e. the same word or sentence may have different importance in a different context. Moreover, current transformer-based document pretraining models suffer from input length constraints. Also, input length becomes a problem for text-rich documents or multi-page documents. (2) documents are

2. Documents are more than words



Figure 3: For (a) we show the samples from RVL-CDIP. The boxes in orange color are grouped OCR bounding boxes. For (b) we plot the accuracies on 16 classes achieved by different models that are trained on the same data.

Effect of visual backbone. Additionally, we apply the trained visual backbone to document object detection on PubLayNet. The performance of the F-RCNN on the validation set is depicted in Table 4. To better compare, we establish two F-RCNN models with: (1) backbone initialized with ResNet-50 pretrained on ImageNet; (2) backbone initialized from UniDoc's pretrained visual backbone. It can be seen that our pretrained backbone outperforms ImageNet-pretrained backbones. By leveraging UniDoc, we can train different variants of the visual backbone and apply them to document-specific downstream applications, without relying on incompatible pretrained backbones from other domains (e.g., natural image). Moreover, the visual backbone of UniDoc does not require any custom layers, and thus any ConvNet architecture can be used in place of ResNet.

5 Conclusion, Limitations, and Future Works

We develop UniDoc, a unified pretraining framework for document understanding. Our model introduces a novel joint training framework that effectively exploits the visual and textual information during pretraining and finetuning. We evaluate the UniDoc comprehensively on three downstream tasks: form understanding, receipt understanding, and document image classification. Extensive empirical analysis demonstrates that the pretraining procedure can take advantage of multimodal inputs and effectively aggregating and aligning visual and textual information of document images with the proxy tasks. This work has a broader impact on document applications. By finetuning the pretrained UniDoc on task-specific data, document processing systems can provide better results and reduce the expensive data annotations costs. In terms of negative social impact, the document images used for pretraining may contain sensitive information and therefore the models trained on such data may inappropriately leak some private information. To address the privacy leakage, it is worthwhile to explore the combination of privacy-preserving learning and self-supervised learning. There are interesting short- and long-term research directions for UniDoc: (1) we freeze the sentence encoder during pretraining and fine-tuning phases due to computational constraints. A better document representation can be learned by jointly training the sentence encoder, visual backbone and cross-attention encoder in a completely end-to-end fashion. (2) Although impressive performance has been achieved in document entity recognition tasks such as form and receipt understanding, the classification accuracy on semi-structured documents such as forms is still inferior to that of rich-text documents. It is possible to devise a better method to model the spatial relationship among words. (3) An interesting direction is to extend UniDoc to multi-page/multilingual document pretraining. Additionally, there exist many text-based labeled document datasets in the NLP domain, such as document summarization. Can we transfer the knowledge learned from the text-based document domain to the image-based document domain? Lastly, in addition to scanned documents, using digital PDF as part of the pretraining data is a direction worth exploring since it provides rich metadata which could be beneficial for multimodal learning.

3. Documents have spatial layout

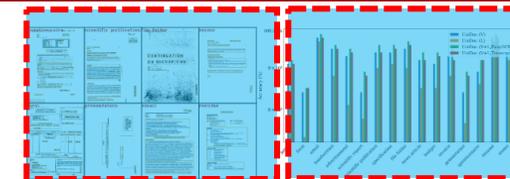


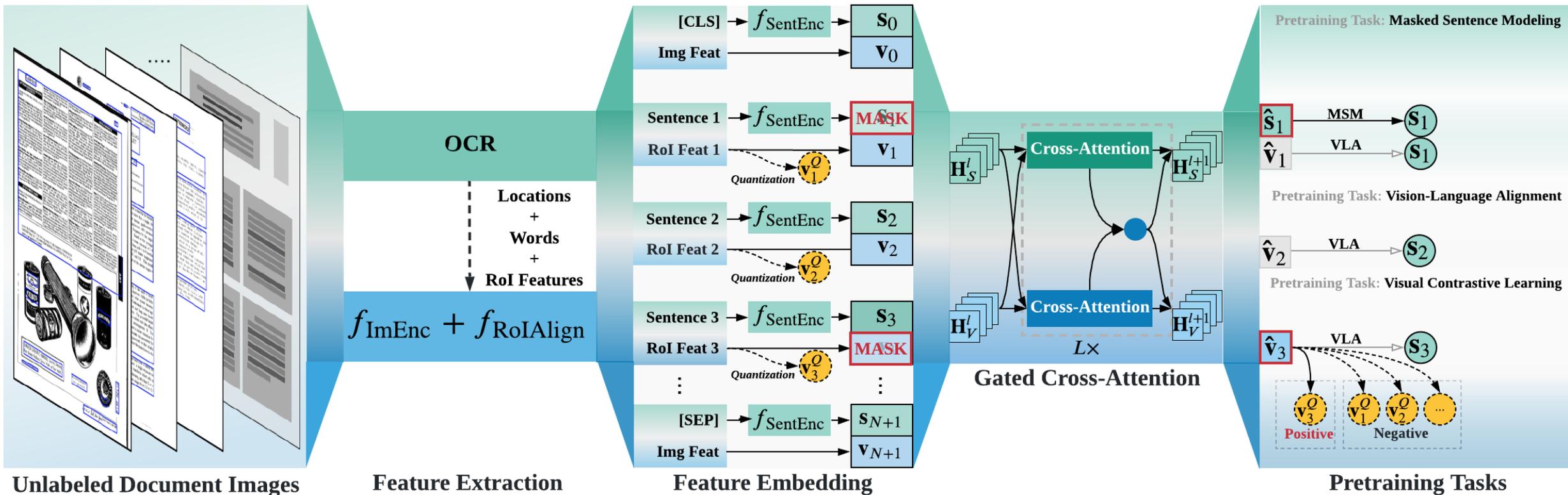
Figure 3: For (a) we show the samples from RVL-CDIP. The boxes in orange color are grouped OCR bounding boxes. For (b) we plot the accuracies on 16 classes achieved by different models that are trained on the same data.

Effect of visual backbone. Additionally, we apply the trained visual backbone to document object detection on PubLayNet. The performance of the F-RCNN on the validation set is depicted in Table 4. To better compare, we establish two F-RCNN models with: (1) backbone initialized with ResNet-50 pretrained on ImageNet; (2) backbone initialized from UniDoc's pretrained visual backbone. It can be seen that our pretrained backbone outperforms ImageNet-pretrained backbones. By leveraging UniDoc, we can train different variants of the visual backbone and apply them to document-specific downstream applications, without relying on incompatible pretrained backbones from other domains (e.g., natural image). Moreover, the visual backbone of UniDoc does not require any custom layers, and thus any ConvNet architecture can be used in place of ResNet.

5 Conclusion, Limitations, and Future Works

We develop UniDoc, a unified pretraining framework for document understanding. Our model introduces a novel joint training framework that effectively exploits the visual and textual information during pretraining and finetuning. We evaluate the UniDoc comprehensively on three downstream tasks: form understanding, receipt understanding, and document image classification. Extensive empirical analysis demonstrates that the pretraining procedure can take advantage of multimodal inputs and effectively aggregating and aligning visual and textual information of document images with the proxy tasks. This work has a broader impact on document applications. By finetuning the pretrained UniDoc on task-specific data, document processing systems can provide better results and reduce the expensive data annotations costs. In terms of negative social impact, the document images used for pretraining may contain sensitive information and therefore the models trained on such data may inappropriately leak some private information. To address the privacy leakage, it is worthwhile to explore the combination of privacy-preserving learning and self-supervised learning. There are interesting short- and long-term research directions for UniDoc: (1) we freeze the sentence encoder during pretraining and fine-tuning phases due to computational constraints. A better document representation can be learned by jointly training the sentence encoder, visual backbone and cross-attention encoder in a completely end-to-end fashion. (2) Although impressive performance has been achieved in document entity recognition tasks such as form and receipt understanding, the classification accuracy on semi-structured documents such as forms is still inferior to that of rich-text documents. It is possible to devise a better method to model the spatial relationship among words. (3) An interesting direction is to extend UniDoc to multi-page/multilingual document pretraining. Additionally, there exist many text-based labeled document datasets in the NLP domain, such as document summarization. Can we transfer the knowledge learned from the text-based document domain to the image-based document domain? Lastly, in addition to scanned documents, using digital PDF as part of the pretraining data is a direction worth exploring since it provides rich metadata which could be beneficial for multimodal learning.

Framework

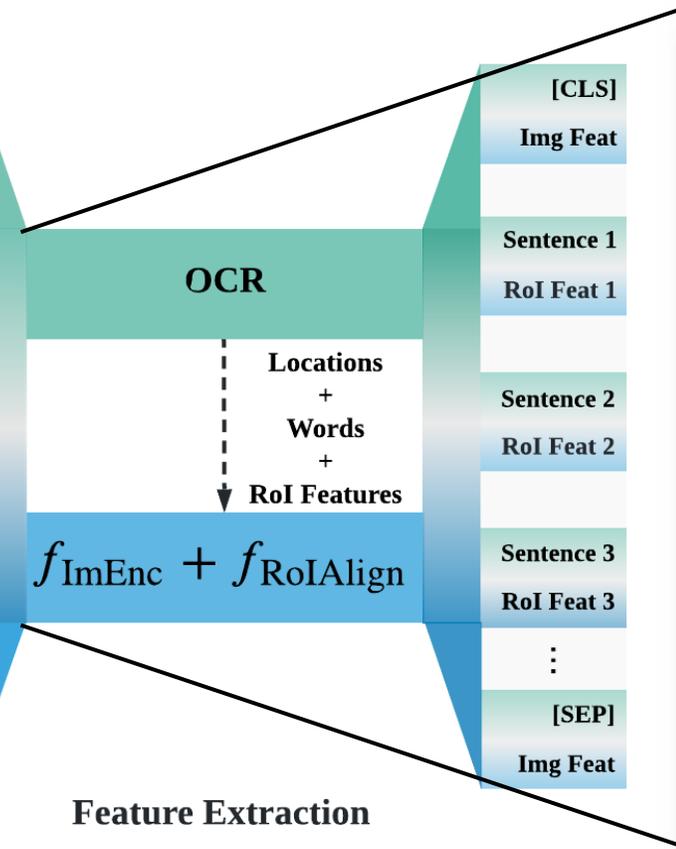


Feature Extraction



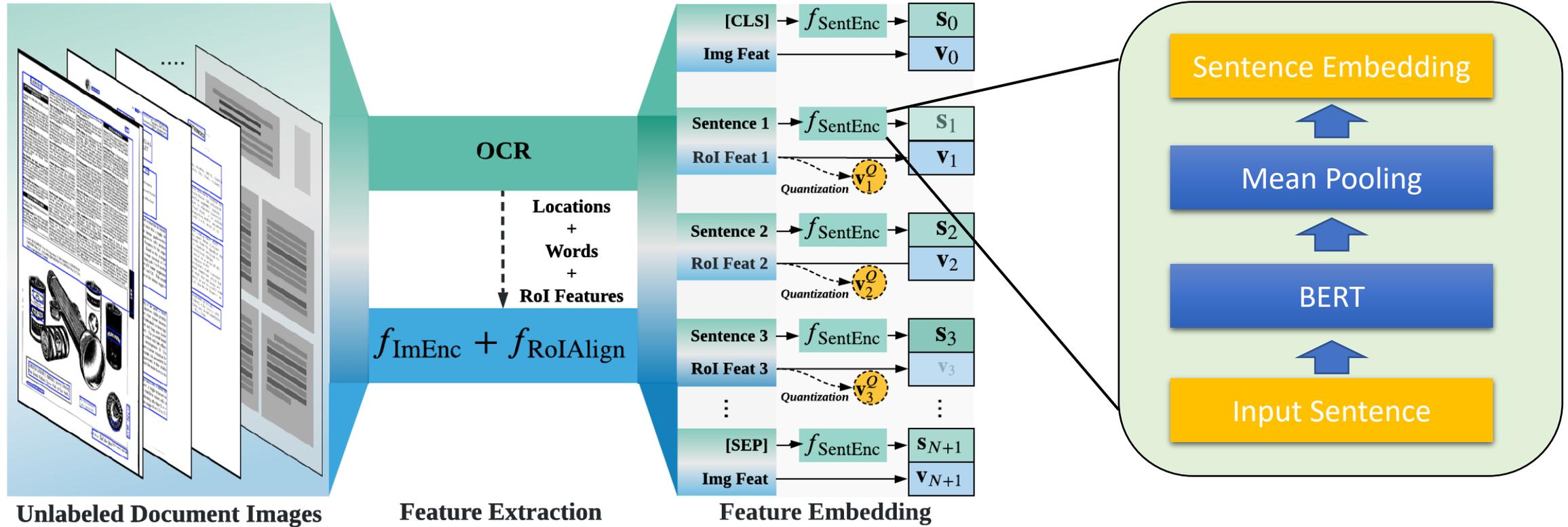
Unlabeled Document Images

Feature Extraction

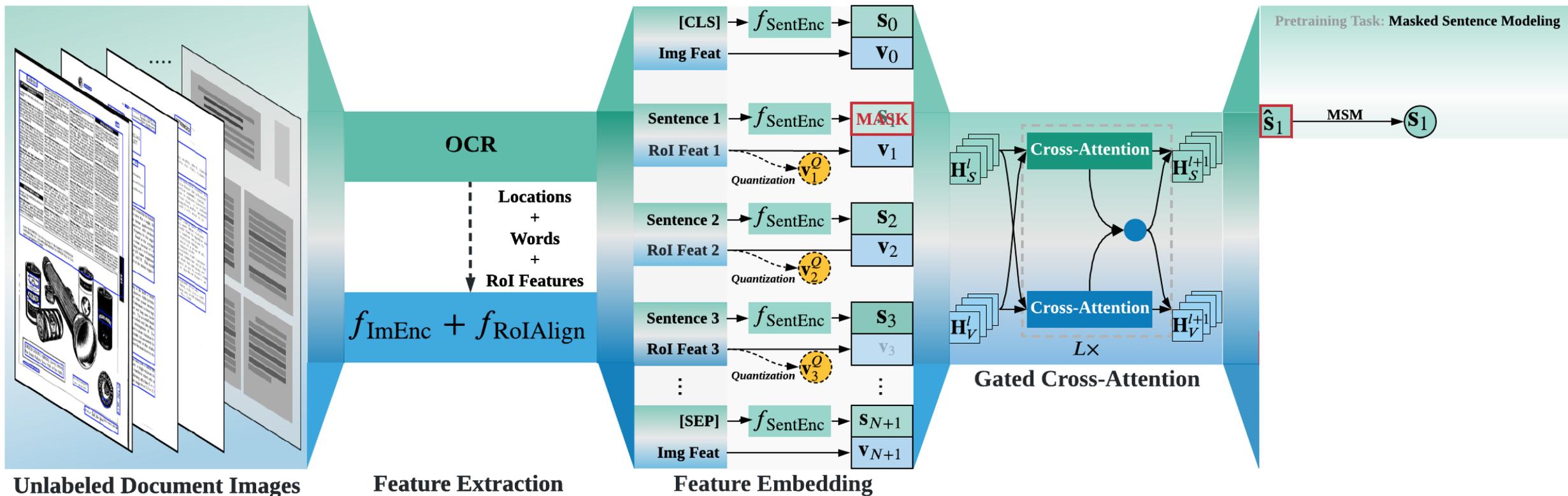


<p>(Paragraph=False)</p>	<p>(Paragraph=False)</p>	<p>(Paragraph=False)</p>	<p>(Paragraph=False)</p>
<p>(Paragraph=True)</p>	<p>(Paragraph=True)</p>	<p>(Paragraph=True)</p>	<p>(Paragraph=True)</p>

Feature Embedding

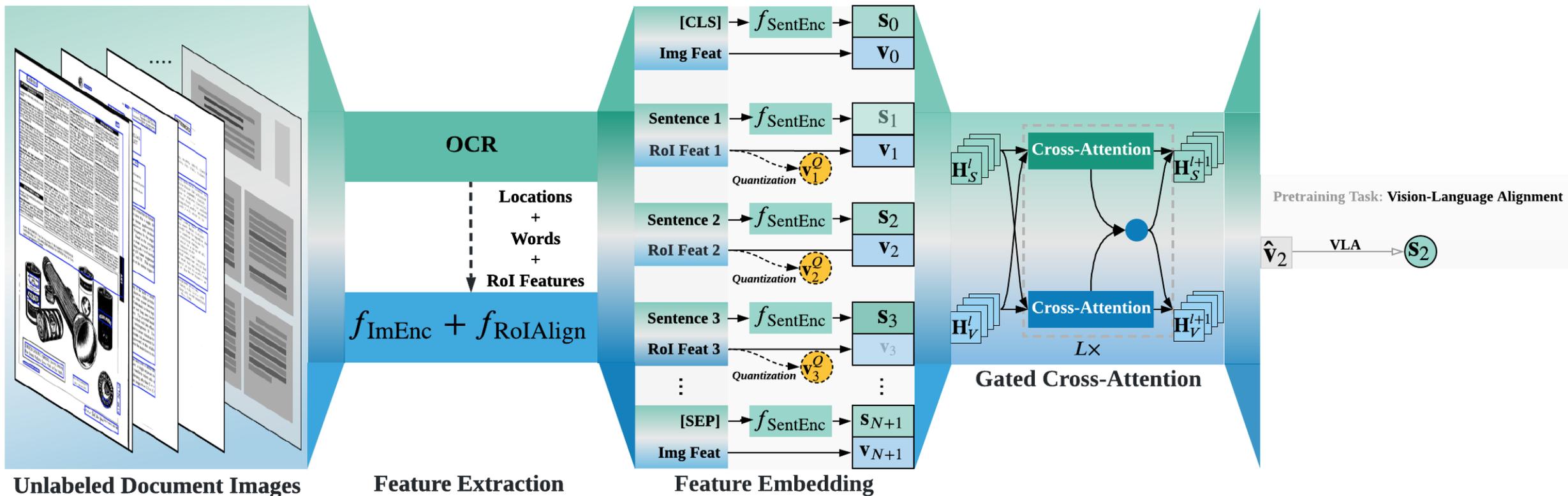


Pretraining Task: *Masked Sentence Modeling*



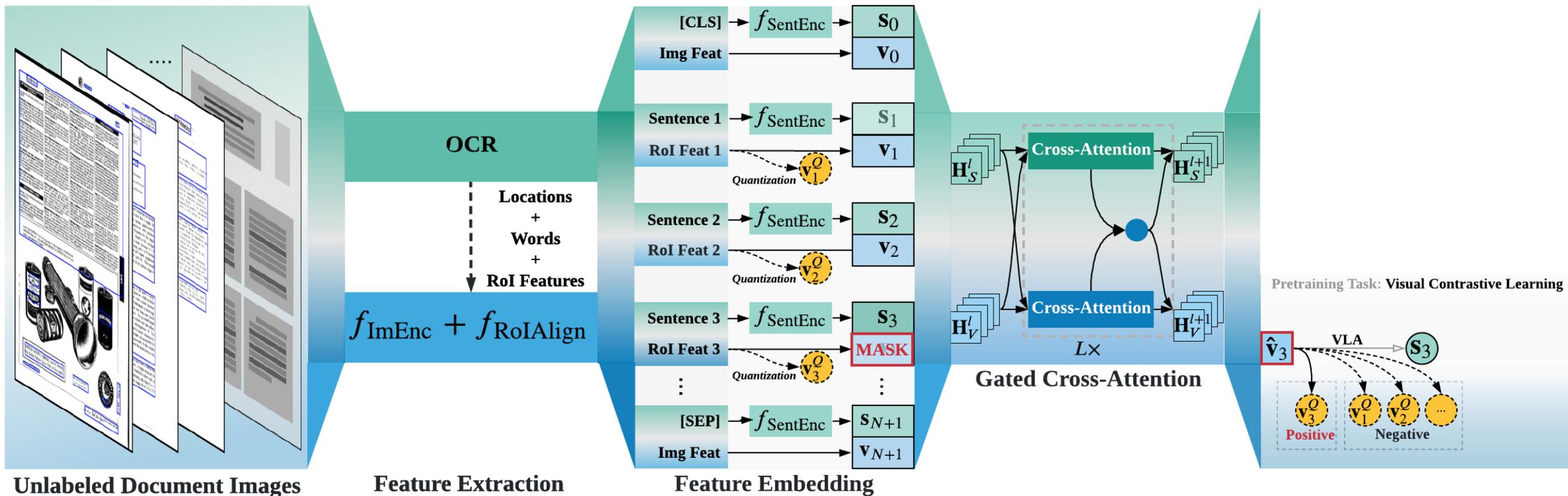
$$\mathcal{L}_{\text{MSM}}(\Theta) = \sum_i \text{smooth}_{L_1}(s_i - f_{\text{UniDoc}}(s_i | s_{\setminus i}, \tilde{\mathbf{V}}))$$

Pretraining Task: *Vision-Language Alignment*



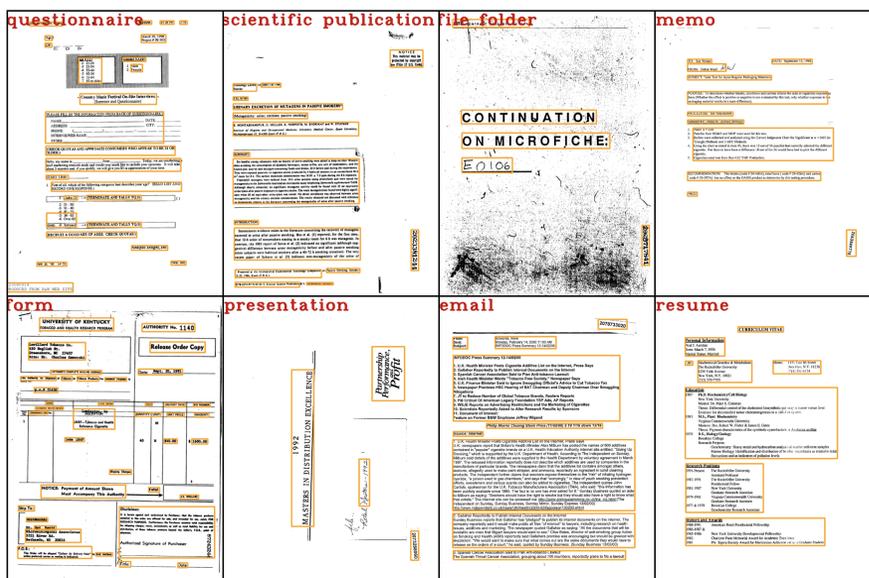
$$\mathcal{L}_{\text{VLA}}(\Theta) = \frac{1}{N \times N} \|f_{\text{Norm}}(\mathbf{S} \cdot \mathbf{S}^\top) - f_{\text{Norm}}(\mathbf{H}_V^L \cdot \mathbf{H}_V^{L\top})\|_F^2$$

Pretraining Task: *Visual Contrastive Learning*



$$\mathcal{L}_{\text{VCL}}(\Theta) = - \sum_{\hat{v}_i \in \tilde{V}} \left(\log \frac{\exp(\text{sim}(\hat{v}_i, v_i^Q)/\kappa)}{\sum_j \exp(\text{sim}(\hat{v}_i, v_j^Q)/\kappa)} \right) + \lambda \frac{1}{CE} \sum_{c=1}^C \sum_{e=1}^E p_{c,e} \log p_{c,e}$$

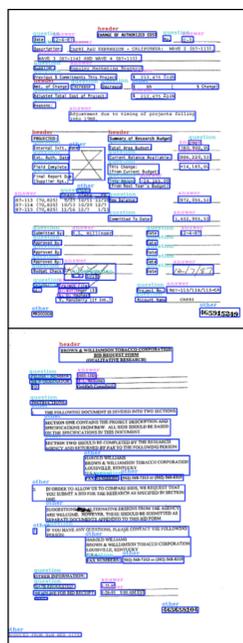
Finetuning Task: *Document Classification*



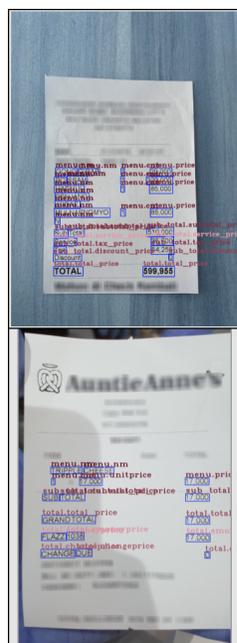
Samples from RVL-CDIP

Method	Pretraining						FUNSD CORP		RVL-CDIP
	Source	#Data	Scale	Max #Words	Modality	#Param.	F1	F1	Accuracy
BERT _{BASE} [5]	–	–	Word	512	L	110M	60.26	89.68	89.81
BERT _{LARGE} [5]	–	–	Word	512	L	340M	65.63	90.25	89.92
LayoutLM _{BASE} [5]	IIT-CDIP	11M	Word	512	L	113M	78.66	94.72	94.42
LayoutLM _{LARGE} [5]	IIT-CDIP	11M	Word	512	L	343M	78.95	94.95	94.43
LayoutLMv2 _{BASE} [5]	IIT-CDIP	11M	Word	512	V+L	200M	82.76	94.95	95.25
LayoutLMv2 _{LARGE} [5]	IIT-CDIP	11M	Word	512	V+L	426M	84.20	96.01	95.64
SelfDoc [6]	RVL-CDIP	320K	Region	50×512	V+L	–	83.36	–	92.81
SelfDoc+VGG-16 [6]	RVL-CDIP	320K	Region	50×512	V+L	–	–	–	93.81
TILT-Base [34]	RVL-CDIP+	1.1M	Word	512	V+L	230M	–	95.11	95.25
TILT-Large [34]	RVL-CDIP+	1.1M	Word	512	V+L	780M	–	96.35	95.52
UDoc	IIT-CDIP	1M	Region	64×512	V+L	272M	87.96	98.85	93.96
UDoc*	IIT-CDIP	1M	Region	64×512	V+L	272M	87.93	98.94	95.05 [‡]

Finetuning Task: *Document Entity Recognition*



FUNSD



CORD

Method	Pretraining						FUNSD CORD		RVL-CDIP
	Source	#Data	Scale	Max #Words	Modality	#Param.	F1	F1	Accuracy
BERT _{BASE} [5]	–	–	Word	512	L	110M	60.26	89.68	89.81
BERT _{LARGE} [5]	–	–	Word	512	L	340M	65.63	90.25	89.92
LayoutLM _{BASE} [5]	IIT-CDIP	11M	Word	512	L	113M	78.66	94.72	94.42
LayoutLM _{LARGE} [5]	IIT-CDIP	11M	Word	512	L	343M	78.95	94.93	94.43
LayoutLMv2 _{BASE} [5]	IIT-CDIP	11M	Word	512	V+L	200M	82.76	94.95	95.25
LayoutLMv2 _{LARGE} [5]	IIT-CDIP	11M	Word	512	V+L	426M	84.20	96.01	95.64
SelfDoc [6]	RVL-CDIP	320K	Region	50×512	V+L	–	83.36	–	92.81
SelfDoc+VGG-16 [6]	RVL-CDIP	320K	Region	50×512	V+L	–	–	–	93.81
TILT-Base [34]	RVL-CDIP+	1.1M	Word	512	V+L	230M	–	95.11	95.25
TILT-Large [34]	RVL-CDIP+	1.1M	Word	512	V+L	780M	–	96.33	95.52
UDoc	IIT-CDIP	1M	Region	64×512	V+L	272M	87.96	98.85	93.96
UDoc*	IIT-CDIP	1M	Region	64×512	V+L	272M	87.93	98.94	95.05 [‡]

At the Poster:

- **Additional details**
- **Quantitative results**
- **Discussion**